Electronic Version
Stylesheet Version v1.1.1

Description

BIOINFORMATICALLY DETECTABLE GROUP OF NOVEL REGULATORY BACTERIAL AND BACTERIAL ASSOCIATED OLIGONUCLEOTIDES AND USES THEREOF

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application is a continuation in part of U.S Patent Application Serial No.10/708,951, filed 2-Apr-04, entitled "Bioinformatically Detectable Group of Novel Regulatory Bacterial and Bacterial Associated Oligonucleotides and Uses Thereof ", the disclosure of which is hereby incorporated by reference and claims priority therefrom; This application also is a continuation in part of U.S. Provisional Patent Application Serial No. 60/521,433 filed 26-Apr-04, entitled "A Microarray for the Detection of MicroRNA Oligonucleotides", the disclosure of which is hereby incorporated by reference and claims priority therefrom.

REFERENCES CITED

[0002] Altschul,S.F., Gish,W., Miller,W., Myers,E.W., and Lipman,D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

[0003] Dan Gusfield, Algorithms on strings, trees, and sequences: computer science and computational biology, Cambridge University Press, 1997.

[0004] Elbashir,S.M., Lendeckel,W., and Tuschl,T. (2001). RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Dev. 15, 188-200.

[0005] Gussow,D. and Clackson,T. (1989). Direct clone characterization from plaques and colonies by the polymerase chain reaction. Nucleic Acids Res. 17, 4000.

[0006] Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D and McKusick VA.(2002).Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.Nucleic Acids Res. 30: 52-55.

[0007] Jenuth,J.P. (2000). The NCBI. Publicly available tools and resources on the Web. Methods Mol. Biol. 132, 301-312.

[0008] Kirkness,E.F. and Kerlavage,A.R. (1997). The TIGR human cDNA database. Methods Mol. Biol. 69, 261-268.

[0009] Krichevsky,A.M., King,K.S., Donahue,C.P., Khrapko,K., and Kosik,K.S. (2003). A microRNA array reveals extensive

regulation of microRNAs during brain development. RNA. 9, 1274-1281.

[0010] Lagos-Quintana,M., Rauhut,R., Lendeckel,W., and Tuschl,T. (2001). Identification of novel genes coding for small expressed RNAs. Science 294, 853-858.

[0011] Lau,N.C., Lim,L.P., Weinstein,E.G., and Bartel,D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science 294, 858-862.

[0012] Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H., and Brown,E.L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. 14, 1675-1680.

[0013] Mathews,D.H., Sabina,J., Zuker,M., and Turner,D.H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 288, 911-940.

[0014] Reinhart,B.J., Slack,F.J., Basson,M., Pasquinelli,A.E., Bettinger,J.C., Rougvie,A.E., Horvitz,H.R., and Ruvkun,G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 403, 901-906.

[0015] Southern,E.M. (1992). Detection of specific sequences

among DNA fragments separated by gel electrophoresis. 1975. Biotechnology 24, 122-139.

[0016] Tom M. Mitchell, Machine Learning, McGraw Hill, 1997.

[0017] Wightman,B., Ha,I., and Ruvkun,G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell 75, 855-862.

BACKGROUND OF THE INVENTION

FIELD OF THE INVENTION

[0018] The present invention relates to a group of bioinformatically detectable novel bacterial oligonucleotides and to a group of bioinformatically detectable novel human oligonucleotides associated with bacterial infections, both are identified here as "Genomic Address Messenger" (GAM) oligonucleotides.

[0019] All of abovementioned oligonucleotides are believed to be related to the microRNA (miRNA) group of oligonucleotides.

DESCRIPTION OF PRIOR ART

[0020] miRNA oligonucleotides are short ~22 nucleotide (nt)-long, non-coding, regulatory RNA oligonucleotides that are found in a wide range of species. miRNA oligonu-

cleotides are believed to function as specific gene translation repressors and are sometimes involved in cell differentiation.

[0021] The ability to detect novel miRNA oligonucleotides is limited by the methodologies used to detect such oligonucleotides. All miRNA oligonucleotides identified so far either present a visibly discernable whole body phenotype, as do Lin-4 and Let-7 (Wightman,B., Ha,I., and Ruvkun,G., Cell 75: 855-862 (1993); Reinhart et al. Nature 403: 901-906 (2000)), or produce sufficient quantities of RNA so as to be detected by standard molecular biological techniques.

[0022] Ninety-three miRNA oligonucleotides have been discovered in several species (Lau et al., Science 294: 858-862 (2001), Lagos-Quintana et al., Science 294: 853-858 (2001)) by sequencing a limited number of clones (300 by Lau and 100 by Lagos-Quintana) of size-fractionated small segments of RNA. miRNAs that were detected in these studies therefore represent the more prevalent among the miRNA oligonucleotide family and cannot be much rarer than 1% of all small ~20 nt-long RNA oligonucleotides.

[0023] The aforementioned studies provide no basis for the de-

tection of miRNA oligonucleotides which either do not present a visually discernable whole body phenotype, or are rare (e.g. rarer than 0.1% of all of the size-fractionated, ~20 nt-long RNA segments that were expressed in the tissues examined), and therefore do not produce large enough quantities of RNA to be detected by standard biological techniques.

[0024] To date, miRNA oligonucleotides have not been detected in bacteria.

[0025] The following U.S. Patents relate to bioinformatic detection of genes: U.S Patent No. 348935, entitled "Statistical algorithms for folding and target accessibility prediction and design of nucleic acids", U.S Patent No. 6,369,195, entitled "Prostate-specific gene for diagnosis, prognosis and management of prostate cancer", and U.S Patent No.6,291,666 entitled "Spike tissue-specific promoter", each of which is hereby incorporated by reference herein.

BRIEF DESCRIPTION OF SEQUENCE LISTING, TABLES AND COMPUTER PROGRAM LISTING

[0026] A sequence listing is attached to the present invention, comprising 4,254,670 genomic sequences, is contained in a file named SEQ_LIST.txt (720288KB, 18-May-04), and is hereby incorporated by reference herein.

[0027] Tables relating to genomic sequences are attached to the present application, appear in the following files (size, creation date) included on CD, incorporated herein: TABLE_1.txt (28.3 MB, 18-May-04), TABLE_2.txt (350 MB, 18-May-04), TABLE_3.txt (5.64 MB, 18-May-04), TABLE_4.txt (17.1 MB, 18-May-04), TABLE_5.txt (5.04 MB, 18-May-04), TABLE_6.txt (536 MB, 18-May-04), TABLE_7_A.txt (619 MB, 18-May-04), TABLE_7_B.txt (340 MB, 18-May-04), TABLE_8_A.txt (619 MB, 18-May-04), TABLE_8_B.txt (619 MB, 18-May-04), TABLE_8_C.txt (619 MB, 18-May-04), TABLE_8_D.txt (457 MB, 18-May-04), TABLE_9.txt (654 MB, 18-May-04), TABLE_10.txt (49.1 MB, 18-May-04), and TABLE_11.txt (79.8 MB, 18-May-04), all of which are incorporated by reference herein. Further, additional tables relating to genomic sequences are attached to the present application, appear in the following files (size, creation date) attached to the application, incorporated herein: TABLE_12.txt (41.1 KB, 18-May-04) and TABLE_13.txt (46.9 KB,18-May-04), are incorporated by reference herein.

[0028] A computer program listing constructed and operative in accordance with a preferred embodiment of the present invention is enclosed on an electronic medium in com-

puter readable form, and is hereby incorporated by reference herein. The computer program listing is contained in 7 files, the name, sizes and creation date of which are as follows: AUXILARY_FILES.txt (117K, 14-Nov-03); EDIT_DISTANCE.txt (144K, 24-Nov-03); FIRST-K.txt (96K, 24-Nov-03); HAIRPIN_PREDICTION.txt (19K, 25-Mar-04); TWO_PHASED_SIDE_SELECTOR.txt (4K, 14-Nov-03); TWO_PHASED_PREDICTOR.txt (74K, 14-Nov-03), and BS_CODE.txt (118K,11-May-04).

SUMMARY OF THE INVENTION

[0029] The present invention relates to a novel group of 3,873 bioinformatically detectable bacterial regulatory RNA oligonucleotides, which repress expression of human target genes, by means of complementary hybridization to binding sites in untranslated regions of these target genes. It is believed that this novel group of bacterial oligonucleotides represents a pervasive bacterial mechanism of attacking a host, and therefore knowledge of this novel group of bacterial oligonucleotides may be useful in preventing and treating bacterial diseases.

[0030] Additionally, the present invention relates to a novel group of 4,363 bioinformatically detectable human regulatory RNA oligonucleotides, which repress expression of

human target genes associated with the bacterial infection, by means of complementary hybridization to binding sites in untranslated regions of these target genes. It is believed that this novel group of human oligonucleotides represents a pervasive novel host response mechanism, and therefore knowledge of this novel group of human oligonucleotides may be useful in preventing and treating bacterial diseases.

[0031] Furthermore, the present invention relates to a novel group of 24,160 bioinformatically detectable bacterial regulatory RNA oligonucleotides, which repress expression of bacterial target genes, by means of complementary hybridization to binding sites in untranslated regions of these bacterial target genes. It is believed that this novel group of bacterial oligonucleotides represents a pervasive novel internal bacterial regulation mechanism, and therefore knowledge of this novel group of bacterial oligonucleotides may be useful in preventing and treating bacterial diseases.

[0032] In addition, the present invention relates to a novel group of 6,100 bioinformatically detectable human regulatory RNA oligonucleotides, which repress expression of bacterial target genes, by means of complementary hybridiza-

tion to binding sites in untranslated regions of these bacterial target genes. It is believed that this novel group of human oligonucleotides represents a pervasive novel antibacterial host defense mechanism, and therefore knowledge of this novel group of human oligonucleotides may be useful in preventing and treating bacterial diseases.

[0033] Also disclosed are 6,056 novel microRNA-cluster like bacterial polynucleotides and 430 novel microRNA-cluster like human polynucleotides, both referred to here as Genomic Record (GR) polynucleotides.

[0034] In various preferred embodiments, the present invention seeks to provide improved method and system for detection and prevention of bacterial diseases, which are mediated by this group of novel oligonucleotides.

[0035] Accordingly, the invention provides several substantially pure nucleic acids (e.g., genomic DNA, cDNA or synthetic DNA) each comprising a novel GAM oligonucleotide, vectors comprising the DNAs, probes comprising the DNAs, a method and system for selectively modulating translation of known target genes utilizing the vectors, and a method and system utilizing the GAM probes to modulate expression of GAM target genes.

[0036] The present invention represents a scientific break-

through, disclosing novel miRNA-like oligonucleotides the number of which is dramatically larger than previously believed existed. Prior-art studies reporting miRNA oligonucleotides ((Lau et al., Science 294:858-862 (2001), Lagos-Quintana et al., Science 294: 853-858 (2001)) discovered 93 miRNA oligonucleotides in several species, including 21 in human, using conventional molecular biology methods, such as cloning and sequencing.

[0037] Molecular biology methodologies employed by these studies are limited in their ability to detect rare miRNA oligonucleotides, since these studies relied on sequencing of a limited number of clones (300 clones by Lau and 100 clones by Lagos-Quintana) of small segments (i.e. size-fractionated) of RNA. miRNA oligonucleotides detected in these studies therefore, represent the more prevalent among the miRNA oligonucleotide family, and are typically not be much rarer than 1% of all small ~20 nt-long RNA oligonucleotides present in the tissue from the RNA was extracted.

[0038] Recent studies state the number of miRNA oligonucleotides to be limited, and describe the limited sensitivity of available methods for detection of miRNA oligonucleotides: "The estimate of 255 human miRNA oligonu-

cleotides is an upper bound implying that no more than 40 miRNA oligonucleotides remain to be identified in mammals" (Lim et al., Science, 299:1540 (2003)); "Estimates place the total number of vertebrate miRNA genes at about 200-250" (Ambros et al. Curr. Biol. 13:807-818 (2003)); and "Confirmation of very low abundance miRNAs awaits the application of detection methods more sensitive than Northern blots" (Ambros et al. Curr. Biol. 13:807-818 (2003)).

[0039] The oligonucleotides of the present invention represent a revolutionary new dimension of genomics and of biology: a dimension comprising a huge number of non-protein-coding oligonucleotides which modulate expression of thousands of proteins and are associated with numerous major diseases. This new dimension disclosed by the present invention dismantles a central dogma that has dominated life-sciences during the past 50 years, a dogma which has emphasized the importance of protein-coding regions of the genome, holding non-protein-coding regions to be of little consequence, often dubbing them "junk DNA".

[0040] Indeed, only in November, 2003 has this long held belief as to the low importance of non-protein-coding regions

been vocally challenged. As an example, an article titled "The Unseen Genome - Gems in the Junk" (Gibbs, W.W. Sci. Am. 289:46-53 (2003)) asserts that the failure to recognize the importance of non-protein- coding regions "may well go down as one of the biggest mistakes in the history of molecular biology." Gibbs further asserts that "what was damned as junk because it was not understood, may in fact turn out to be the very basis of human complexity." The present invention provides a dramatic leap in understanding specific important roles of non-protein-coding regions.

[0041] An additional scientific breakthrough of the present invention is a novel conceptual model disclosed by the present invention, which conceptual model is preferably used to encode in a genome the determination of cell differentiation, utilizing oligonucleotides and polynucleotides of the present invention.

[0042] Using the bioinformatic engine of the present invention, 21,916 bacterial GAM oligonucleotides and their respective precursors and targets have been detected and 6,100 human GAM oligonucleotides and their respective precursors and targets have been detected. These bioinformatic predictions are supported by robust biological studies.

Microarray experiments validated expression of 346 of the human GAM oligonucleotides of the present invention. Of these, 311 received an extremely high score: over six standard deviations higher than the background "noise" of the microarray, and over two standard deviations above their individual "mismatch" control probes and 33 received a high score: over four standard deviations higher than the background "noise" of the microarray. Further, 38 GAM oligonucleotides were sequenced.

[0043] In various preferred embodiments, the present invention seeks to provide an improved method and system for specific modulation of the expression of specific target genes involved in significant human diseases. It also provides an improved method and system for detection of the expression of novel oligonucleotides of the present invention, which modulate these target genes. In many cases, the target genes may be known and fully characterized, however in alternative embodiments of the present invention, unknown or less well characterized genes may be targeted.

[0044] A "Nucleic acid" is defined as a ribonucleic acid (RNA) molecule, or a deoxyribonucleic acid (DNA) molecule, or complementary deoxyribonucleic acid (cDNA), comprising

either naturally occurring nucleotides or non-naturally occurring nucleotides.

[0045] "Substantially pure nucleic acid", "Isolated Nucleic Acid", "Isolated Oligoucleotide" and "Isolated Polynucleotide" are defined as a nucleic acid that is free of the genome of the organism from which the nucleic acid is derived, and include, for example, a recombinant nucleic acid which is incorporated into a vector, into an autonomously replicating plasmid or virus, or into the genomic nucleic acid of a prokaryote or eukaryote at a site other than its natural site; or which exists as a separate molecule (e.g., a cDNA or a genomic or cDNA fragment produced by PCR or restriction endonuclease digestion) independent of other nucleic acids.

[0046] An "Oligonucleotide" is defined as a nucleic acid comprising 2-139 nts, or preferably 16-120 nts. A "Polynucleotide" is defined as a nucleic acid comprising 140-5000 nts, or preferably 140-1000 nts.

[0047] A "Complementary" sequence is defined as a first nucleotide sequence which reverses complementary of a second nucleotide sequence: the first nucleotide sequence is reversed relative to a second nucleotide sequence, and wherein each nucleotide in the first nucleotide sequence is

complementary to a corresponding nucleotide in the second nucleotide sequence (e.g. ATGGC is the complementary sequence of GCCAT).

[0048] "Hybridization", "Binding" and "Annealing" are defined as hybridization, under in vivo physiological conditions, of a first nucleic acid to a second nucleic acid, which second nucleic acid is at least partially complementary to the first nucleic acid.

[0049] A "Hairpin Structure" is defined as an oligonucleotide having a nucleotide sequence that is 50-140 nts in length, the first half of which nucleotide sequence is at least partially complementary to the second part thereof, thereby causing the nucleic acid to fold onto itself, forming a secondary hairpin structure.

[0050] A "Hairpin-Shaped Precursor" is defined as a Hairpin Structure which is processed by a Dicer enzyme complex, yielding an oligonucleotide which is about 19 to about 24 nts in length.

[0051] "Inhibiting translation" is defined as the ability to prevent synthesis of a specific protein encoded by a respective gene by means of inhibiting the translation of the mRNA of this gene. For example, inhibiting translation may include the following steps: (1) a DNA segment encodes an

RNA, the first half of whose sequence is partially complementary to the second half thereof; (2) the precursor folds onto itself forming a hairpin-shaped precursor; (3) a Dicer enzyme complex cuts the hairpin-shaped precursor yielding an oligonucleotide that is approximately 22 nt in length; (4) the oligonucleotide binds complementarily to at least one binding site, having a nucleotide sequence that is at least partially complementary to the oligonucleotide, which binding site is located in the mRNA of a target gene, preferably in the untranslated region (UTR) of a target gene, such that the binding inhibits translation of the target protein.

[0052] A "Translation inhibitor site" is defined as the minimal nucleotide sequence sufficient to inhibit translation.

[0053] The present invention describes novel GAM oligonucleotides, detected using a bioinformatic engine described hereinabove. The ability of this detection engine has been demonstrated using stringent algorithmic criteria, showing that the engine has both high sensitivity, indicated by the high detection rate of published miRNA oligonucleotides and their targets, as well as high specificity, indicated by the low amount of "background" hairpin candidates passing its filters. Laboratory tests, based both on

sequencing of predicted GAM oligonucleotides and on microarray experiments, validated 381 of the GAM oligonucleotides in the present invention. Further, almost all of the bacterial target genes (6,141 of the 7,351) and almost all of the human target genes (64 out of 76) described in the present invention are bound by one or more of the 381 human GAM oligonucleotides validated by the microarray experiments.

[0054] There is thus provided in accordance with a preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which is endogenously processed from a hairpin-shaped precursor, and anneals to a portion of a mRNA transcript of a target gene, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs: 1-385 and 386-49787.

[0055] There is further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide having a nucleotide sequence selected from the group consisting of SEQ ID NOs: 1-385 and 386-49787.

[0056] There is still further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable first oligonucleotide which is a portion of a mRNA transcript of a target gene, and anneals to a second oligonucleotide that is endogenously processed from a hairpin precursor, wherein binding of the first oligonucleotide to the second oligonucleotide represses expression of the target gene, and wherein nucleotide sequence of the second nucleotide is selected from the group consisting of SEQ ID NOs: 1-385 and 386-49787.

[0057] There is additionally provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable oligonucleotide having a nucleotide sequence selected from the group consisting of SEQ ID NOs: 2337129-4223628.

[0058] There is moreover provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Bordetella pertussis infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a

nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 2.

[0059] There is further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Brucella suis 1330 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 3.

[0060] There is still further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Chlamydia trachomatis infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 4.

[0061] There is additionally provided in accordance with another

preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Chlamydophila pneumoniae AR39 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 5.

[0062] There is moreover provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Chlamydophila pneumoniae CWL029 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 6.

[0063] There is further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which an-

neals to a portion of a mRNA transcript of a target gene associated with Chlamydophila pneumoniae J138 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 7.

[0064] There is still further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Chlamydophila pneumoniae TW-183 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 8.

[0065] There is additionally provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Coxiella burnetii RSA 493 infection,

wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 9.

[0066] There is moreover provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Escherichia coli CFT073 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 10.

[0067] There is further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Haemophilus influenzae Rd infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence

identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 11.

[0068] There is still further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Leptospira interrogans serovar lai str. 56601 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 12.

[0069] There is additionally provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Listeria monocytogenes EGD-e infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the

group consisting of SEQ ID NOs shown in Table 13 row 13.

[0070] There is moreover provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Mycobacterium avium subsp. paratuberculosis infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 14.

[0071] There is further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Mycobacterium bovis subsp bovis AF2122/97 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown

in Table 13 row 15.

[0072] There is still further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Mycobacterium leprae infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 16.

[0073] There is additionally provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Mycobacterium tuberculosis CDC1551 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 17.

[0074] There is moreover provided in accordance with another

preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Mycobacterium tuberculosis H37Rv infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 18.

[0075] There is further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Neisseria meningitidis MC58 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 19.

[0076] There is still further provided in accordance with another preferred embodiment of the present invention a bioinfor-

matically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Neisseria meningitidis Z2491 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 20.

[0077] There is additionally provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Pseudomonas aeruginosa PA01 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 21.

[0078] There is moreover provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which an-

neals to a portion of a mRNA transcript of a target gene associated with Pseudomonas putida KT2440 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 22.

[0079] There is further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Rickettsia prowazekii infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 23.

[0080] There is still further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Salmonella enterica enterica serovar Typhi

infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 24.

[0081] There is additionally provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Salmonella enterica enterica serovar Typhi Ty2 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 25.

[0082] There is moreover provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Salmonella typhimurium LT2 infection, wherein binding of the oligonucleotide to the mRNA tran-

script represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 26.

[0083] There is further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Shigella flexneri 2a str. 2457T infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 27.

[0084] There is still further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Shigella flexneri 2a str. 301 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and

wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 28.

[0085] There is additionally provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Staphylococcus aureus subsp. aureus Mu50 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 29.

[0086] There is moreover provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Staphylococcus aureus subsp. aureus MW2 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80%

sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 30.

[0087] There is further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Staphylococcus aureus subsp. aureus N315 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 31.

[0088] There is still further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Streptococcus pneumoniae R6 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the

group consisting of SEQ ID NOs shown in Table 13 row 32.

[0089] There is additionally provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Streptococcus pneumoniae TIGR4 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 33.

[0090] There is moreover provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Streptococcus pyogenes M1 GAS infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row

34.

[0091] There is further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Streptococcus pyogenes MGAS315 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 35.

[0092] There is still further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Streptococcus pyogenes MGAS8232 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 36.

[0093] There is additionally provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Streptococcus pyogenes SSI-1 infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 37.

[0094] There is moreover provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Treponema pallidum subsp. pallidum str. Nichols infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 38.

[0095] There is further provided in accordance with another pre-

ferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Yersinia pestis infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 39.

[0096] There is still further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which anneals to a portion of a mRNA transcript of a target gene associated with Yersinia pestis KIM infection, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide has at least 80% sequence identity with a nucleotide sequence selected from the group consisting of SEQ ID NOs shown in Table 13 row 40.

[0097] There is additionally provided in accordance with another preferred embodiment of the present invention a method for treatment of a disease involving a tissue in which a protein is pathologically expressed to an undesirable ex-

tent, the protein having a messenger RNA, the method including: providing a material which modulates activity of a microRNA oligonucleotide which binds complementarily to a segment of the messenger RNA, and introducing the material into the tissue, causing modulation of the activity of the microRNA oligonucleotide and thereby modulating expression of the protein in a desired manner.

[0098] There is moreover provided in accordance with another preferred embodiment of the present invention a method for treatment of a disease involving tissue in which a protein is pathologically expressed to an undesirable extent, the protein having a messenger RNA, the method including: providing a material which at least partially binds a segment of the messenger RNA that is bound complementarily by a microRNA oligonucleotide, thereby modulating expression of the protein, and introducing the material into the tissue, thereby modulating expression of the protein.

[0099] There is further provided in accordance with another preferred embodiment of the present invention a method for treatment of a disease involving a tissue in which a protein is pathologically over-expressed, the protein having a messenger RNA, the method including: providing a mi-

croRNA oligonucleotide which binds complementarily to a segment of the messenger RNA, and introducing the microRNA oligonucleotide into the tissue, causing the microRNA oligonucleotide to bind complementarily to a segment of the messenger RNA and thereby inhibit expression of the protein.

[0100] There is still further provided in accordance with another preferred embodiment of the present invention a method for treatment of a disease involving a tissue in which a protein is pathologically over-expressed, the protein having a messenger RNA, the method including: providing a chemically-modified microRNA oligonucleotide which binds complementarily to a segment of the messenger RNA, and introducing the chemically-modified microRNA oligonucleotide into the tissue, causing the microRNA oligonucleotide to bind complementarily to a segment of the messenger RNA and thereby inhibit expression of the protein.

[0101] There is additionally provided in accordance with another preferred embodiment of the present invention a method for treatment of a disease involving a tissue in which a protein is pathologically under-expressed, the protein having a messenger RNA, the method including: providing

an oligonucleotide that inhibits activity of a microRNA oligonucleotide which binds complementarily to a segment of the messenger RNA, and introducing the oligonucleotide into the tissue, causing inhibition of the activity of the microRNA oligonucleotide and thereby promotion of translation of the protein.

[0102] There is moreover provided in accordance with another preferred embodiment of the present invention a method for treatment of a disease involving a tissue in which a protein is pathologically under-expressed, the protein having a messenger RNA, the method including: providing a chemically-modified oligonucleotide that inhibits activity of a microRNA oligonucleotide which binds complementarily to a segment of the messenger RNA, and introducing the chemically-modified oligonucleotide into the tissue, causing inhibition of the activity of the microRNA oligonucleotide and thereby promotion of translation of the protein.

[0103] There is further provided in accordance with another preferred embodiment of the present invention a method for diagnosis of a disease involving a tissue in which a protein is expressed to abnormal extent, the protein having a messenger RNA, the method including: assaying a mi-

croRNA oligonucleotide which at least partially binds a segment of the messenger RNA and modulates expression of the protein, thereby providing an indication of at least one parameter of the disease.

[0104] There is still further provided in accordance with another preferred embodiment of the present invention a method for detection of expression of an oligonucleotide, the method including: determining a first nucleotide sequence of a first oligonucleotide, which first nucleotide sequence is not complementary to a genome of an organism, receiving a second nucleotide sequence of a second oligonucleotide whose expression is sought to be detected, designing a third nucleotide sequence that is complementary to the second nucleotide sequence of the second oligonucleotide, and a fourth nucleotide sequence that is complementary to a fifth nucleotide sequence which is different from the second nucleotide sequence of the second oligonucleotide by at least one nucleotide, synthesizing a first oligonucleotide probe having a sixth nucleotide sequence including the third nucleotide sequence followed by the first nucleotide sequence of the first oligonucleotide, and a second oligonucleotide probe having a seventh nucleotide sequence including the fourth

nucleotide sequence followed by the first nucleotide sequence of the first oligonucleotide, locating the first oligonucleotide probe and the second oligonucleotide probe on a microarray platform, receiving an RNA test sample from at least one tissue of the organism, obtaining size-fractionated RNA from the RNA test sample, amplifying the size-fractionated RNA, hybridizing the adaptor-linked RNA with the first and second oligonucleotide probes on the microarray platform, and determining expression of the first oligonucleotide in the at least one tissue of the organism, based at least in part on the hybridizing.

[0105] There is additionally provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated polynucleotide which is endogenously processed into a plurality of hairpin-shaped precursor oligonucleotides, each of which is endogenously processed into a respective oligonucleotide, which in turn anneals to a portion of a mRNA transcript of a target gene, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene.

[0106] There is moreover provided in accordance with another preferred embodiment of the present invention a bioinfor-

matically detectable isolated oligonucleotide which is endogenously processed from a hairpin-shaped precursor, and anneals to a portion of a mRNA transcript of a target gene, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the target gene does not encode a protein.

[0107] There is further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which is endogenously processed from a hairpin-shaped precursor, and anneals to a portion of a mRNA transcript of a target gene, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein a function of the oligonucleotide includes modulation of cell type.

[0108] There is still further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable isolated oligonucleotide which is endogenously processed from a hairpin-shaped precursor, and anneals to a portion of a mRNA transcript of a target gene, wherein binding of the oligonucleotide to the mRNA transcript represses expression of the target gene, and wherein the oligonucleotide is maternally transferred by a

cell to at least one daughter cell of the cell, and a function of the oligonucleotide includes modulation of cell type of the daughter cell.

[0109] There is additionally provided in accordance with another preferred embodiment of the present invention a method for bioinformatic detection of microRNA oligonucleotides, the method including: bioinformatically detecting a hairpin-shaped precursor oligonucleotide, bioinformatically detecting an oligonucleotide which is endogenously processed from the hairpin-shaped precursor oligonucleotide, and bioinformatically detecting a target gene of the oligonucleotide wherein the oligonucleotide anneals to at least one portion of a mRNA transcript of the target gene, and wherein the binding represses expression of the target gene, and the target gene is associated with a disease.

Brief Description of Drawings

[0110] Fig. 1 is a simplified diagram illustrating a mode by which an oligonucleotide of a novel group of oligonucleotides of the present invention modulates expression of known target genes;

[0111] Fig. 2 is a simplified block diagram illustrating a bioinformatic oligonucleotide detection system capable of detect-

ing oligonucleotides of the novel group of oligonucleotides of the present invention, which system is constructed and operative in accordance with a preferred embodiment of the present invention;

[0112] Fig. 3 is a simplified flowchart illustrating operation of a mechanism for training of a computer system to recognize the novel oligonucleotides of the present invention, which mechanism is constructed and operative in accordance with a preferred embodiment of the present invention;

[0113] Fig. 4A is a simplified block diagram of a non-coding genomic sequence detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0114] Fig. 4B is a simplified flowchart illustrating operation of a non-coding genomic sequence detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0115] Fig. 5A is a simplified block diagram of a hairpin detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0116] Fig. 5B is a simplified flowchart illustrating operation of a hairpin detector constructed and operative in accordance

with a preferred embodiment of the present invention;

[0117] Fig. 6A is a simplified block diagram of a Dicer-cut location detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0118] Fig. 6B is a simplified flowchart illustrating training of a Dicer-cut location detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0119] Fig. 6C is a simplified flowchart illustrating operation of a Dicer-cut location detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0120] Fig. 7A is a simplified block diagram of a target gene binding site detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0121] Fig. 7B is a simplified flowchart illustrating operation of a target gene binding site detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0122] Fig. 8 is a simplified flowchart illustrating operation of a function and utility analyzer constructed and operative in accordance with a preferred embodiment of the present

invention;

[0123] Fig. 9 is a simplified diagram describing a novel bioinformatically-detected group of regulatory polynucleotides, referred to here as Genomic Record (GR) polynucleotides, each of which encodes an "operon-like" cluster of novel microRNA-like oligonucleotides, which in turn modulate expression of one or more target genes;

[0124] Fig. 10 is a block diagram illustrating different utilities of novel oligonucleotides and novel operon-like polynucleotides, both of the present invention;

[0125] Figs. 11A and 11B are simplified diagrams which, when taken together, illustrate a mode of oligonucleotide therapy applicable to novel oligonucleotides of the present invention;

[0126] Fig. 12A is a bar graph illustrating performance results of a hairpin detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0127] Fig. 12B is a line graph illustrating accuracy of a Dicer-cut location detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0128] Fig. 12C is a bar graph illustrating performance results of the target gene binding site detector 118, constructed and

operative in accordance with a preferred embodiment of the present invention.

[0129] Fig. 13 is a summary table of laboratory results validating expression of novel human oligonucleotides detected by a bioinformatic oligonucleotide detection engine constructed and operative in accordance with a preferred embodiment of the present invention, thereby validating its efficacy;

[0130] Fig. 14A is a schematic representation of an "operon-like" cluster of novel human hairpin sequences detected by a bioinformatic oligonucleotide detection engine constructed and operative in accordance with a preferred embodiment of the present invention, and non-GAM hairpin sequences used as negative controls thereto;

[0131] Fig. 14B is a schematic representation of secondary folding of hairpins of the operon-like cluster of Fig. 14A;

[0132] Fig. 14C is a picture of laboratory results demonstrating expression of novel oligonucleotides of Figs. 14A and 14B and lack of expression of the negative controls, thereby validating efficacy of bioinformatic detection of GAM oligonucleotides and GR polynucleotides detected by a bioinformatic oligonucleotide detection engine, constructed and operative in accordance with a preferred em-

bodiment of the present invention;

[0133] Fig. 15A is an annotated sequence of EST72223 comprising known human microRNA oligonucleotide MIR98 and novel human oligonucleotide GAM25 PRECURSOR detected by the oligonucleotide detection system of the present invention; and

[0134] Figs. 15B, 15C and 15D are pictures of laboratory results demonstrating laboratory confirmation of expression of known human oligonucleotide MIR98 and of novel bioinformatically-detected human GAM25 RNA respectively, both of Fig. 15A, thus validating the bioinformatic oligonucleotide detection system of the present invention;

[0135] Fig. 16A, 16B and 16C are schematic diagrams which, when taken together, represent methods of designing primers to identify specific hairpin oligonucleotides in accordance with a preferred embodiment of the present invention.

[0136] Fig. 17A is a simplified flowchart illustrating construction of a microarray constructed and operative to identify novel oligonucleotides of the present invention, in accordance with a preferred embodiment of the present invention;

[0137] Fig. 17B is a simplified block diagram illustrating design of a microarray constructed and operative to identify novel

oligonucleotides of the present invention, in accordance with a preferred embodiment of the present invention;

[0138] Fig. 17C is a flowchart illustrating a mode of preparation and amplification of a cDNA library in accordance with a preferred embodiment of the present invention;

[0139] Fig. 18A is a line graph showing results of detection of known microRNA oligonucleotides and of novel GAM oligonucleotides, using a microarray constructed and operative in accordance with a preferred embodiment of the present invention;

[0140] Fig. 18B is a line graph showing specificity of hybridization of a microarray constructed and operative in accordance with a preferred embodiment of the present invention; and

[0141] Fig. 18C is a summary table demonstrating detection of known microRNA oligonucleotides using a microarray constructed and operative in accordance with a preferred embodiment of the present invention.

BRIEF DESCRIPTION OF SEQUENCES

[0142] A Sequence Listing of genomic sequences of the present invention designated SEQ ID NO:1 through SEQ ID: 4,254,670 is attached to this application, and is hereby incorporated herein. The genomic listing comprises the

following nucleotide sequences: nucleotide sequences of 21,916 bacterial and 6,100 human GAM precursors of respective novel oligonucleotides of the present invention; nucleotide sequences of 32,713 bacterial and 11,428 human GAM RNA oligonucleotides of respective novel oligonucleotides of the present invention; and nucleotide sequences of 1,507,219 target gene binding sites of respective novel oligonucleotides of the present invention.

Detailed Description

[0143] Reference is now made to Fig. 1, which is a simplified diagram describing a plurality of novel bioinformatically-detected oligonucleotide of the present invention referred to here as the Genomic Address Messenger (GAM) oligonucleotide, which modulates expression of respective target genes whose function and utility are known in the art.

[0144] GAM is a novel bioinformatically detectable regulatory, non-protein-coding, miRNA-like oligonucleotide. The method by which GAM is detected is described with additional reference to Figs. 1-8.

[0145] The GAM PRECURSOR is preferably encoded by a bacterial genome. Alternatively or additionally, the GAM PRECURSOR is preferably encoded by the human genome. The GAM TARGET GENE is a gene encoded by the human

genome. Alternatively or additionally, the GAM TARGET GENE is a gene encoded by a bacterial genome.

[0146] The GAM PRECURSOR encodes a GAM PRECURSOR RNA. Similar to other miRNA oligonucleotides, the GAM PRECURSOR RNA does not encode a protein.

[0147] GAM PRECURSOR RNA folds onto itself, forming GAM FOLDED PRECURSOR RNA, which has a two-dimensional "hairpin" structure. GAM PRECURSOR RNA folds onto itself, forming GAM FOLDED PRECURSOR RNA, which has a two-dimensional "hairpin structure". As is well-known in the art, this "hairpin structure" is typical of RNA encoded by known miRNA precursor oligonucleotides and is due to the full or partial complementarity of the nucleotide sequence of the first half of an miRNA precursor to theRNA that is encoded by a miRNA oligonucleotide to the nucleotide sequence of the second half thereof.

[0148] A complementary sequence is a sequence which is reversed and wherein each nucleotide is replaced by a complementary nucleotide, as is well known in the art (e.g. ATGGC is the complementary sequence of GCCAT).

[0149] An enzyme complex designated DICER COMPLEX, an enzyme complex composed of Dicer RNaseIII together with other necessary proteins, cuts the GAM FOLDED PRECUR-

SOR RNA yielding a single-stranded ~22 nt-long RNA segment designated GAM RNA.

[0150] GAM TARGET GENE encodes a corresponding messenger RNA, designated GAM TARGET RNA. As is typical of mRNA of a protein-coding gene, each GAM TARGET RNAs of the present invention comprises three regions, as is typical of mRNA of a protein-coding gene: a 5' untranslated region, a protein-coding region and a 3' untranslated region, designated 5'UTR, PROTEIN-CODING and 3'UTR, respectively.

[0151] GAM RNA binds complementarily to one or more target binding sites located in the untranslated regions of each of the GAM TARGET RNAs of the present invention. This complementary binding is due to the partial or full complementarity between the nucleotide sequence of GAM RNA and the nucleotide sequence of each of the target binding sites. As an illustration, Fig. 1 shows three such target binding sites, designated BINDING SITE I, BINDING SITE II and BINDING SITE III, respectively. It is appreciated that the number of target binding sites shown in Fig. 1 is only illustrative and that any suitable number of target binding sites may be present. It is further appreciated that although Fig. 1 shows target binding sites only in the

3'UTR region, these target binding sites may instead be located in the 5'UTR region or in both the 3'UTR and 5'UTR regions.

[0152] The complementary binding of GAM RNA to target binding sites on GAM TARGET RNA, such as BINDING SITE I, BINDING SITE II and BINDING SITE III, inhibits the translation of each of the GAM TARGET RNAs of the present invention into repsective GAM TARGET PROTEIN, shown surrounded by a broken line.

[0153] It is appreciated that the GAM TARGET GENE in fact represents a plurality of GAM target genes. The mRNA of each one of this plurality of GAM target genes comprises one or more target binding sites, each having a nucleotide sequence which is at least partly complementary to GAM RNA and which when bound by GAM RNA causes inhibition of translation of the GAM target mRNA into a corresponding GAM target protein.

[0154] The mechanism of the translational inhibition that is exerted by GAM RNA on one or more GAM TARGET GENEs may be similar or identical to the known mechanism of translational inhibition exerted by known miRNA oligonucleotides.

[0155] The nucleotide sequences of each of a plurality of GAM

oligonucleotides that are described by Fig. 1 and their respective genomic sources and genomic locations are set forth in Tables 1-3, hereby incorporated herein.

[0156] The nucleotide sequences of GAM PRECURSOR RNAs, and a schematic representation of a predicted secondary folding of GAM FOLDED PRECURSOR RNAs, of each of a plurality of GAM oligonucleotides that are described by Fig. 1 are set forth in Table 4, hereby incorporated herein.

[0157] The nucleotide sequences of "diced" GAM RNAs of each of a plurality of GAM oligonucleotides that are described by Fig. 1 are set forth in Table 5, hereby incorporated herein.

[0158] The nucleotide sequences of target binding sites, such as BINDING SITE I, BINDING SITE II and BINDING SITE III that are found on GAM TARGET RNAs of each of a plurality of GAM oligonucleotides that are described by Fig. 1, and a schematic representation of the complementarity of each of these target binding sites to each of a plurality of GAM RNAs that are described by Fig. 1 are set forth in Tables 6-7, hereby incorporated herein.

[0159] It is appreciated that the specific functions and accordingly the utilities of each of a plurality of GAM oligonucleotides that are described by Fig. 1 are correlated with and may be deduced from the identity of the GAM TARGET

GENES inhibited thereby, and whose functions are set forth in Table 8, hereby incorporated herein.

[0160] Studies documenting the well known correlations between each of a plurality of GAM TARGET GENEs that are described by Fig. 1 and the known gene functions and related diseases are listed in Table 9, hereby incorporated herein.

[0161] The present invention discloses a novel group of bacterial and human oligonucleotides, belonging to the miRNA-like oligonucleotide group, here termed GAM oligonucleotides, for which a specific complementary binding has been determined bioinformatically.

[0162] Reference is now made to Fig. 2, which is a simplified block diagram illustrating a bioinformatic oligonucleotide detection system and method constructed and operative in accordance with a preferred embodiment of the present invention.

[0163] An important feature of the present invention is a bioinformatic oligonucleotide detection engine 100, which is capable of bioinformatically detecting oligonucleotides of the present invention.

[0164] The functionality of the bioinformatic oligonucleotide detection engine 100 includes receiving expressed RNA data

102, sequenced DNA data 104, and protein function data 106; performing a complex process of analysis of this data as elaborated hereinbelow, and based on this analysis provides information, designated by reference numeral 108, identifying and describing features of novel oligonucleotides.

[0165] Expressed RNA data 102 comprises published expressed sequence tags (EST) data, published mRNA data, as well as other published RNA data. Sequenced DNA data 104 comprises alphanumeric data representing genomic sequences and preferably including annotations such as information indicating the location of known protein-coding regions relative to the genomic sequences.

[0166] Protein function data 106 comprises information from scientific publications e.g. physiological functions of known proteins and their connection, involvement and possible utility in treatment and diagnosis of various diseases.

[0167] Expressed RNA data 102 and sequenced DNA data 104 may preferably be obtained from data published by the National Center for Biotechnology Informatiion (NCBI) at the National Institute of Health (NIH) (Jenuth,J.P. (2000). Methods Mol. Biol. 132:301-312(2000), herein incorporated by reference) as well as from various other pub-

lished data sources. Protein function data 106 may preferably be obtained from any one of numerous relevant published data sources, such as the Online Mendelian Inherited Disease In Man (OMIM(TM), Hamosh et al., Nucleic Acids Res. 30: 52-55(2002)) database developed by John Hopkins University, and also published by NCBI (2000).

[0168] Prior to or during actual detection of bioinformatically-detected group of novel oligonucleotides 108 by the bioinformatic oligonucleotide detection engine 100, bioinformatic oligonucleotide detection engine training & validation functionality 110 is operative. This functionality uses one or more known miRNA oligonucleotides as a training set to train the bioinformatic oligonucleotide detection engine 100 to bioinformatically recognize miRNA-like oligonucleotides, and their respective potential target binding sites. Bioinformatic oligonucleotide detection engine training & validation functionality 110 is further described hereinbelow with reference to Fig. 3.

[0169] The bioinformatic oligonucleotide detection engine 100 preferably comprises several modules which are preferably activated sequentially, and are described as follows:

[0170] A non-protein-coding genomic sequence detector 112 operative to bioinformatically detect non-protein-coding

genomic sequences. The non-protein-coding genomic sequence detector 112 is further described herein below with reference to Figs. 4A and 4B.

[0171] A hairpin detector 114 operative to bioinformatically detect genomic "hairpin-shaped" sequences, similar to GAM FOLDED PRECURSOR RNA (Fig. 1). The hairpin detector 114 is further described herein below with reference to Figs. 5A and 5B.

[0172] A Dicer-cut location detector 116 operative to bioinformatically detect the location on a GAM FOLDED PRECURSOR RNA which is enzymatically cut by DICER COMPLEX (Fig. 1), yielding "diced" GAM RNA. The Dicer-cut location detector 116 is further described herein below with reference to Figs. 6A-6C.

[0173] A target gene binding site detector 118 operative to bioinformatically detect target genes having binding sites, the nucleotide sequence of which is partially complementary to that of a given genomic sequence, such as a nucleotide sequence cut by DICER COMPLEX. The target gene binding site detector 118 is further described hereinbelow with reference to Figs. 7A and 7B.

[0174] A function & utility analyzer, designated by reference numeral 120, is operative to analyze the function and utility

of target genes in order to identify target genes which have a significant clinical function and utility. The function & utility analyzer 120 is further described hereinbelow with reference to Fig. 8

[0175] According to an embodiment of the present invention, the bioinformatic oligonucleotide detection engine 100 may employ a cluster of 40 personal computers (PCs; XEON (R), 2.8GHz, with 80GB storage each) connected by Ethernet to eight servers (2-CPU, XEON (TM) 1.2-2.2GHz, with ~200GB storage each) and combined with an 8-processor server (8-CPU, Xeon 550Mhz w/ 8GB RAM) connected via 2 HBA fiber-channels to an EMC CLARIION (TM) 100-disks, 3.6 Terabyte storage device. A preferred embodiment of the present invention may also preferably comprise software that utilizes a commercial database software program, such as MICROSOFT (TM) SQL Server 2000.

[0176] According to a preferred embodiment of the present invention, the bioinformatic oligonucleotide detection engine 100 may employ a cluster of 80 Servers (XEON (R), 2.8GHz, with 80GB storage each) connected by Ethernet to eight servers (2-CPU, XEON (TM) 1.2-2.2GHz, with ~200GB storage each) and combined with storage device

(Promise Technology Inc., RM8000) connected to an 8-disks, 2 Terabytes total. A preferred embodiment of the present invention may also preferably comprise software that utilizes a commercial database software program, such as MICROSOFT (TM) SQL Server 2000. It is appreciated that the abovementioned hardware configuration is not meant to be limiting and is given as an illustration only. The present invention may be implemented in a wide variety of hardware and software configurations.

[0177] The present invention discloses 21,916 bacterial and 6,100 human novel oligonucleotides of the GAM group of oligonucleotides, which have been detected bioinformatically and 6,056 bacterial and 430 novel polynucleotides of the GR group of polynucleotides, which have been detected bioinformatically. Laboratory confirmation of bioinformatically predicted oligonucleotides of the GAM group of oligonucleotides, and several bioinformatically predicted polynucleotides of the GR group of polynucleotides, is described hereinbelow with reference to Figs. 13-15D, Fig. 18 and Table 12.

[0178] Reference is now made to Fig. 3, which is a simplified flowchart illustrating operation of a preferred embodiment of the bioinformatic oligonucleotide detection engine

training & validation functionality 110 described hereinabove with reference to Fig. 2.

[0179] bioinformatic oligonucleotide detection engine training & validation functionality 110 begins by training the bioinformatic oligonucleotide detection engine 100 (Fig. 2) to recognize one or more known miRNA oligonucleotides, as designated by reference numeral 122. This training step comprises hairpin detector training & validation functionality 124, further described hereinbelow with reference to Fig. 5A, Dicer-cut location detector training & validation functionality 126, further described hereinbelow with reference to Fig. 6A and 6B, and target gene binding site detector training & validation functionality 128, further described hereinbelow with reference to Fig. 7A.

[0180] Next, the bioinformatic oligonucleotide detection engine training & validation functionality 110 is operative bioinformatically detect novel oligonucleotides, using bioinformatic oligonucleotide detection engine 100 (Fig. 2), as designated by reference numeral 130. Wet lab experiments are preferably conducted in order to validate expression and preferably function of some samples of the novel oligonucleotides detected by the bioinformatic oligonucleotide detection engine 100, as designated by

reference numeral 132. Figs. 13A-15D, Fig. 18 and Table 12 illustrate examples of wet lab validation of sample novel human oligonucleotides bioinformatically-detected in accordance with a preferred embodiment of the present invention.

[0181] Reference is now made to Fig. 4A, which is a simplified block diagram of a preferred implementation of the non-protein-coding genomic sequence detector 112 described hereinabove with reference to Fig. 2. The non-protein-coding genomic sequence detector 112 preferably receives at least two types of published genomic data: Expressed RNA data 102 and sequenced DNA data 104. The expressed RNA data 102 may include, inter alia, EST data, EST clusters data, EST genome alignment data and mRNA data. Sources for expressed RNA data 102 include NCBI dbEST, NCBI UniGene clusters and mapping data, and TIGR gene indices (Kirkness F. and Kerlavage, A.R., Methods Mol. Biol. 69:261-268 (1997)). Sequenced DNA data 104 may include sequence data (FASTA format files), and feature annotations (GenBank file format) mainly from NCBI databases. Based on the abovementioned input data, the non-protein-coding genomic sequence detector 112 produces a plurality of non-protein-coding genomic se-

quences 136. Preferred operation of the non-protein-coding genomic sequence detector 112 is described hereinbelow with reference to Fig. 4B.

[0182] Reference is now made to Fig. 4B, which is a simplified flowchart illustrating a preferred operation of the non-protein-coding genomic sequence detector 112 of Fig. 2. Detection of non-protein-coding genomic sequences 136, generally preferably progresses along one of the following two paths:

[0183] A first path for detecting non-protein-coding genomic sequences 136 (Fig. 4A) begins with receipt of a plurality of known RNA sequences, such as EST data. Each RNA sequence is first compared with known protein-coding DNA sequences, in order to select only those RNA sequences which are non-protein-coding, i.e. intergenic or intronic sequences. This can preferably be performed by using one of many alignment algorithms known in the art, such as BLAST (Altschul et al., J. Mol. Biol. 215:403-410 (1990)). This sequence comparison preferably also provides localization of the RNA sequence on the DNA sequences.

[0184] Alternatively, selection of non-protein-coding RNA sequences and their localization on the DNA sequences can be performed by using publicly available EST cluster data

and genomic mapping databases, such as the UNIGENE database published by NCBI or the TIGR database. Such databases, map expressed RNA sequences to DNA sequences encoding them, find the correct orientation of EST sequences, and indicate mapping of ESTs to protein-coding DNA regions, as is well known in the art. Public databases, such as TIGR, may also be used to map an EST to a cluster of ESTs, known in the art as Tentative Human Consensus and assumed to be expressed as one segment. Publicly available genome annotation databases, such as NCBI's GenBank, may also be used to deduce expressed intronic sequences.

[0185] Optionally, an attempt may be made to "expand" the non-protein RNA sequences thus found, by searching for transcription start and end signals, respectively upstream and downstream of the location of the RNA on the DNA, as is well known in the art.

[0186] A second path for detecting non-protein-coding genomic sequences 136 (Fig. 4A) begins with receipt of DNA sequences. The DNA sequences are parsed into non-protein-coding sequences, using published DNA annotation data, by extracting those DNA sequences which are between known protein-coding sequences. Next, tran-

scription start and end signals are sought. If such signals are found, and depending on their robustness, probable expressed non-protein-coding genomic sequences are obtained. Such approach is especially useful for identifying novel GAM oligonucleotides which are found in proximity to other known miRNA oligonucleotides, or other wet lab validated GAM oligonucleotides. Since, as described hereinbelow with reference to Fig. 9, GAM oligonucleotides are frequently found in clusters; sequences located near known miRNA oligonucleotides are more likely to contain novel GAM oligonucleotides. Optionally, sequence orthology, i.e. sequence conservation in an evolutionary related species, may be used to select genomic sequences having a relatively high probability of containing expressed novel GAM oligonucleotides. It is appreciated that in detecting non-human GAM oligonucleotides of the present invention the bioinformatic oligonucleotide detection engine 100 utilizes the input genomic sequences, without filtering protein-coding regions detected by the non-protein-coding genomic sequence detector 112, hence non-protein-coding genomic sequences 136 refers to GENOMIC SEQUENCES only.

[0187] Reference is now made to Fig. 5A, which is a simplified

block diagram of a preferred implementation of the hairpin detector 114 described hereinabove with reference to Fig. 2.

[0188] The goal of the hairpin detector 114 is to detect hairpin-shaped genomic sequences, similar to those of known miRNA oligonucleotides. A hairpin-shaped genomic sequence is a genomic sequence, having a first half which is at least partially complementary to a second half thereof, which causes the halves to folds onto themselves, thereby forming a hairpin structure, as mentioned hereinabove with reference to Fig. 1.

[0189] The hairpin detector 114 (Fig. 2) receives a plurality of non-protein-coding genomic sequences 136 (Fig. 4A). Following operation of hairpin detector training & validation functionality 124 (Fig. 3), the hairpin detector 114 is operative to detect and output hairpin-shaped sequences, which are found in the non-protein-coding genomic sequences 136. The hairpin-shaped sequences detected by the hairpin detector 114 are designated hairpin structures on genomic sequences 138. A preferred mode of operation of the hairpin detector 114 is described hereinbelow with reference to Fig. 5B.

[0190] hairpin detector training & validation functionality 124 in-

cludes an iterative process of applying the hairpin detector 114 to known hairpin-shaped miRNA precursor sequences, calibrating the hairpin detector 114 such that it identifies a training set of known hairpin-shaped miRNA precursor sequences, as well as other similarly hairpin-shaped sequences. In a preferred embodiment of the present invention, the hairpin detector training & validation functionality 124 trains the hairpin detector 114 and validates each of the steps of operation thereof described hereinbelow with reference to Fig. 5B

[0191] The hairpin detector training & validation functionality 124 preferably uses two sets of data: the aforesaid training set of known hairpin-shaped miRNA precursor sequences, such as hairpin-shaped miRNA precursor sequences of 440 miRNA oligonucleotides of H. sapiens, M. musculus, C. elegans, C. Brigssae and D. Melanogaster, annotated in the RFAM database (Griffiths-Jones 2003), and a background set of about 1000 hairpin-shaped sequences found in expressed non-protein-coding human genomic sequences. The background set is expected to comprise some valid, previously undetected hairpin-shaped miRNA-like precursor sequences, and many hairpin-shaped sequences which are not hairpin-shaped

miRNA-like precursors.

[0192] In a preferred embodiment of the present invention the efficacy of the hairpin detector 114 (Fig. 2) is confirmed. For example, when a similarity threshold is chosen such that 87% of the known hairpin-shaped miRNA precursors are successfully predicted, only 21.8% of the 1000 background set of hairpin-shaped sequences are predicted to be hairpin-shaped miRNA-like precursors.

[0193] Reference is now made to Fig. 5B, which is a simplified flowchart illustrating preferred operation of the hairpin detector 114 of Fig. 2. The hairpin detector 114 preferably initially uses a secondary structure folding algorithm based on free-energy minimization, such as the MFOLD algorithm, described in Mathews et al. J. Mol. Biol. 288:911-940 (1999) and Zuker, M. Nucleic Acids Res. 31: 3406-3415 (2003), the disclosure of which is hereby incorporated by reference. This algorithm is operative to calculate probable secondary structure folding patterns of the non-protein-coding genomic sequences 136 (Fig. 4A) as well as the free-energy of each of these probable secondary folding patterns. The secondary structure folding algorithm, such as the MFOLD algorithm (Mathews, 1997; Zuker 2003), typically provides a listing of the base-

pairing of the folded shape, i.e. a listing of each pair of connected nucleotides in the sequence.

[0194] Next, the hairpin detector 114 analyzes the results of the secondary structure folding patterns, in order to determine the presence and location of hairpin folding structures. The goal of this second step is to assess the base-pairing listing provided by the secondary structure folding algorithm, in order to determine whether the base-pairing listing describes one or more hairpin type bonding pattern. Preferably, sequence segment corresponding to a hairpin structure is then separately analyzed by the secondary structure folding algorithm in order to determine its exact folding pattern and free-energy.

[0195] The hairpin detector 114 then assesses the hairpin structures found by the previous step, comparing them to hairpin structures of known miRNA precursors, using various characteristic hairpin structure features such as its free-energy and its thermodynamic stability, the amount and type of mismatched nucleotides and the existence of sequence repeat-elements, number of mismatched nucleotides in positions 18-22 counting from loop, and Percent of G nucleotide. Only hairpins that bear statistically significant resemblance to the training set of hairpin

structures of known miRNA precursors, according to the abovementioned parameters, are accepted.

[0196] In a preferred embodiment of the present invention, similarity to the training set of hairpin structures of known miRNA precursors is determined using a "similarity score" which is calculated using a multiplicity of terms, where each term is a function of one of the abovementioned hairpin structure features. The parameters of each function are found heuristically from the set of hairpin structures of known miRNA precursors, as described hereinabove with reference to hairpin detector training & validation functionality 124 (Fig. 3). The selection of the features and their function parameters is optimized so as to achieve maximized separation between the distribution of similarity scores validated miRNA precursor hairpin structures, and the distribution of similarity scores of hairpin structures detected in the background set mentioned hereinabove with reference to Fig. 5B.

[0197] In an alternative preferred embodiment of the present invention, the step described in the preceding paragraph may be split into two stages. A first stage implements a simplified scoring method, typically based on thresholding a subset of the hairpin structure features described

hereinabove, and may employ a minimum threshold for hairpin structure length and a maximum threshold for free-energy. A second stage is preferably more stringent, and preferably employs a full calculation of the weighted sum of terms described hereinabove. The second stage preferably is performed only on the subset of hairpin structures that survived the first stage.

[0198] The hairpin detector 114 also attempts to select hairpin structures whose thermodynamic stability is similar to that of hairpin structures of known miRNA precursors. This may be achieved in various ways. A preferred embodiment of the present invention utilizes the following methodology, preferably comprising three logical steps:

[0199] First, the hairpin detector 114 attempts to group hairpin structures into "families" of closely related hairpin structures. As is known in the art, a secondary structure folding algorithm typically provides multiple alternative folding patterns, for a given genomic sequence and indicates the free-energy of each alternative folding pattern. It is a particular feature of the present invention that the hairpin detector 114 preferably assesses the various hairpin structures appearing in the various alternative folding patterns and groups' hairpin structures which appear at

identical or similar sequence locations in various alternative folding patterns into common sequence location based "families" of hairpins. For example, all hairpin structures whose center is within 7 nucleotides of each other may be grouped into a "family". Hairpin structures may also be grouped into a "family" if their nucleotide sequences are identical or overlap to a predetermined degree.

[0200] It is also a particular feature of the present invention that the hairpin structure "families" are assessed in order to select only those families which represent hairpin structures that are as thermodynamically stable as those of hairpin structures of known miRNA precursors. Preferably only families which are represented in at least a selected majority of the alternative secondary structure folding patterns, typically 65%, 80% or 100% are considered to be sufficiently stable. Our tests suggest that only about 50% of the hairpin structures, predicted by the MFOLD algorithm with default parameters, are members of sufficiently stable families, comparing to about 90% of the hairpin structures that contain known miRNAs. This percent depends on the size of the fraction that was fold. In an alternative embodiment of the present invention we use frac-

tions of size 1000 nts as preferable size. Different embodiment uses other sizes of genomics sequences, more or less strict demand for representation in the alternative secondary structure folding patterns.

[0201] It is an additional particular feature of the present invention that the most suitable hairpin structure is selected from each selected family. For example, a hairpin structure which has the greatest similarity to the hairpin structures appearing in alternative folding patterns of the family may be preferred. Alternatively or additionally, the hairpin structures having relatively low free-energy may be preferred.

[0202] Alternatively or additionally considerations of homology to hairpin structures of other organisms and the existence of clusters of thermodynamically stable hairpin structures located adjacent to each other along a sequence may be important in selection of hairpin structures. The tightness of the clusters in terms of their location and the occurrence of both homology and clusters may be of significance.

[0203] Reference is now made to Figs. 6A-6C, which together describe the structure and operation of the Dicer-cut location detector 116, described hereinabove with reference

to Fig. 2.

[0204] Reference is now made to Fig. 6A, which is a simplified block diagram of a preferred implementation of the Dicer-cut location detector 116. The goal of the Dicer-cut location detector 116 is to detect the location in which the DICER COMPLEX, described hereinabove with reference to Fig. 1, dices GAM FOLDED PRECURSOR RNA, yielding GAM RNA.

[0205] The Dicer-cut location detector 116 therefore receives a plurality of hairpin structures on genomic sequences, designated by reference numeral 138 (Fig. 5A), and following operation of Dicer-cut location detector training & validation functionality 126 (Fig 3), is operative to detect a plurality of Dicer-cut sequences from hairpin structures, designated by reference numeral 140.

[0206] Reference is now made to Fig. 6B, which is a simplified flowchart illustrating a preferred implementation of Dicer-cut location detector training & validation functionality 126.

[0207] A general goal of the Dicer-cut location detector training & validation functionality 126 is to analyze the Dicer-cut locations of known diced miRNA on respective hairpin-shaped miRNA precursors in order to determine a com-

mon pattern in these locations, which can be used to predict Dicer-cut locations on GAM folded precursor RNAs.

[0208] The Dicer-cut locations of known miRNA precursors are obtained and studied. Locations of the 5' and/or 3' ends of the known diced miRNA oligonucleotides are preferably represented by their respective distances from the 5' end of the corresponding hairpin-shaped miRNA precursor. Additionally or alternatively, the 5' and/or 3' ends of the known diced miRNA oligonucleotides are preferably represented by the relationship between their locations and the locations of one or more nucleotides along the hairpin-shaped miRNA precursor. Additionally or alternatively, the 5' and/or 3' ends of the known diced miRNA oligonucleotides are preferably represented by the relationship between their locations and the locations of one or more bound nucleotide pairs along the hairpin-shaped miRNA precursor. Additionally or alternatively, the 5' and/or 3' ends of the known diced miRNA oligonucleotides are preferably represented by the relationship between their locations and the locations of one or more mismatched nucleotide pairs along the hairpin-shaped miRNA precursor. Additionally or alternatively, the 5' and/or 3' ends of the known diced miRNA oligonucleotides are preferably

represented by the relationship between their locations and the locations of one or more unmatched nucleotides along the hairpin-shaped miRNA precursor. Additionally or alternatively, locations of the 5' and/or 3' ends of the known diced miRNA oligonucleotides are preferably represented by their respective distances from the loop located at the center of the corresponding hairpin-shaped miRNA precursor.

[0209] One or more of the foregoing location metrics may be employed in the Dicer-cut location detector training & validation functionality 126. Additionally, metrics related to the nucleotide content of the diced miRNA and/or of the hairpin-shaped miRNA precursor may be employed.

[0210] In a preferred embodiment of the present invention, Dicer-cut location detector training & validation functionality 126 preferably employs standard machine learning techniques known in the art of machine learning to analyze existing patterns in a given "training set" of examples. Standard machine learning techniques are capable, to a certain degree, of detecting patterns in examples to which they have not been previously exposed that are similar to those in the training set. Such machine learning techniques include, but are not limited to neural net-

works, Bayesian Modeling, Bayesian Networks, Support Vector Machines (SVM), Genetic Algorithms, Markovian Modeling, Maximum Likelihood Modeling, Nearest Neighbor Algorithms, Decision Trees and other techniques, as is well-known in the art.

[0211] In accordance with an embodiment of the present invention, two or more classifiers or predictors based on the abovementioned machine learning techniques are separately trained on the abovementioned training set, and are used jointly in order to predict the Dicer-cut location. As an example, Fig. 6B illustrates operation of two classifiers, a 3' end recognition classifier and a 5' end recognition classifier. Most preferably, the Dicer-cut location detector training & validation functionality 126 implements a "best-of-breed" approach employing a pair of classifiers based on the abovementioned Bayesian Modeling and Nearest Neighbor Algorithms, and accepting only "potential GAM RNAs" that score highly on one of these predictors. In this context, "high scores" means scores that have been demonstrated to have low false positive value when scoring known miRNA oligonucleotides. Alternatively, the Dicer-cut location detector training & validation functionality 126 may implement operation of more or less than

two classifiers.

[0212] Predictors used in a preferred embodiment of the present invention are further described hereinbelow with reference to Fig. 6C. A computer program listing of a computer program implementation of the Dicer-cut location detector training & validation functionality 126 is enclosed on an electronic medium in computer-readable form, and is hereby incorporated by reference herein.

[0213] When evaluated on the abovementioned validation set of 440 published miRNA oligonucleotides using k-fold cross validation (Mitchell, 1997) with k = 3, the performance of the resulting predictors is as follows: In 70% of known miRNA oligonucleotides, a 5' end location is correctly determined by a Support Vector Machine predictor within up to two nucleotides; a Nearest Neighbor (EDIT DISTANCE) predictor achieves 56% accuracy (247/440); and a Two-Phased Predictor that uses Bayesian modeling (TWO PHASED) achieves 80% accuracy (352/440) when only the first phase is used. When the second phase (strand choice) is implemented by a naive Bayesian model, the accuracy is 55% (244/440), and when the K-nearest-neighbor modeling is used for the second phase, 374/440 decisions are made and the accuracy is 65% (242/374). A K-near-

est-neighbor predictor (FIRST-K) achieves 61% accuracy (268/440). The accuracies of all predictors are considerably higher on top-scoring subsets of published miRNA oligonucleotides.

[0214] Finally, in order to validate the efficacy and accuracy of the Dicer-cut location detector 116, a sample of novel oligonucleotides detected thereby is preferably selected, and validated by wet lab experiments. Laboratory results validating the efficacy of the Dicer-cut location detector 116 are described hereinbelow with reference to Figs. 13-15D, Fig. 18 and also in the enclosed file Table 12.

[0215] Reference is now made to Fig. 6C, which is a simplified flowchart illustrating an operation of a Dicer-cut location detector 116 (Fig. 2), constructed and operative in accordance with a preferred embodiment of the present invention. The Dicer-cut location detector 116 preferably comprises a machine learning computer program module, which is trained to recognize Dicer-cut locations on known hairpin-shaped miRNA precursors, and based on this training, is operable to detect Dicer-cut locations of novel GAM RNA (Fig. 1) on GAM FOLDED PRECURSOR RNA (Fig. 1). In a preferred embodiment of the present invention, the Dicer-cut location module preferably utilizes

machine learning algorithms, including but not limited to Support Vector Machine, Bayesian modeling, Nearest Neighbors, and K-nearest-neighbor algorithms that are known in the art.

[0216] When initially assessing a novel GAM FOLDED PRECURSOR RNA, each 19-24 nt-long segment thereof is considered to be a potential GAM RNA, because the Dicer-cut location is initially unknown.

[0217] For each such potential GAM RNA, the location of its 5' end or the locations of its 5' and 3' ends are scored by at least one recognition classifier or predictor, operating on features such as the follwing: Locations of the 5' and/or 3' ends of the known diced miRNA oligonucleotides, which are preferably represented by their respective distances from the 5' end of the corresponding hairpin-shaped miRNA precursor. Additionally or alternatively, the 5' and/or 3' ends of the known diced miRNA oligonucleotides, which are preferably represented by the relationship between their locations and the locations of one or more nucleotides along the hairpin-shaped miRNA precursor. Additionally or alternatively, the 5' and/or 3' ends of the known diced miRNA oligonucleotides, which are preferably represented by the relationship between their loca-

tions and the locations of one or more bound nucleotide pairs along the hairpin-shaped miRNA precursor. Additionally or alternatively, the 5' and/or 3' ends of the known diced miRNA oligonucleotides, which are preferably represented by the relationship between their locations and the locations of one or more mismatched nucleotide pairs along the hairpin-shaped miRNA precursor. Additionally or alternatively, the 5' and/or 3' ends of the known diced miRNA oligonucleotides, which are preferably represented by the relationship between their locations and the locations of one or more unmatched nucleotides along the hairpin-shaped miRNA precursor. Additionally or alternatively, locations of the 5' and/or 3' ends of the known diced miRNA oligonucleotides, which are preferably represented by their respective distances from the loop located at the center of the corresponding hairpin-shaped miRNA precursor. Additionally or alternatively, metrics related to the nucleotide content of the diced miRNA and/or of the hairpin-shaped miRNA precursor.

[0218] In a preferred embodiment of the present invention, the Dicer-cut location detector 116 (Fig. 2) may use a Support Vector Machine predictor.

[0219] In another preferred embodiment of the present invention, the Dicer-cut location detector 116 (Fig. 2) preferably employs an "EDIT DISTANCE" predictor, which seeks sequences that are similar to those of known miRNA oligonucleotides, utilizing a Nearest Neighbor algorithm, where a similarity metric between two sequences is a variant of the Edit Distance algorithm (Gusfield, 1997). The EDIT DISTANCE predictor is based on an observation that miRNA oligonucleotides tend to form clusters, the members of which show marked sequence similarity.

[0220] In yet another preferred embodiment of the present invention, the Dicer-cut location detector 116 (Fig. 2) preferably uses a "TWO PHASE" predictor, which predicts the Dicer-cut location in two distinct phases: (a) selecting a double-stranded segment of the GAM FOLDED PRECURSOR RNA (Fig. 1) comprising the GAM RNA by naive Bayesian modeling and (b) detecting which strand of the double-stranded segment contains GAM RNA (Fig. 1) by employing either naive or K-nearest-neighbor modeling. K-nearest-neighbor modeling is a variant of the "FIRST-K" predictor described hereinbelow, with parameters optimized for this specific task. The "TWO PHASE" predictor may be operated in two modes: either utilizing only the

first phase and thereby producing two alternative Dicer-cut location predictions, or utilizing both phases and thereby producing only one final Dicer-cut location.

[0221] In still another preferred embodiment of the present invention, the Dicer-cut location detector 116 preferably uses a "FIRST-K" predictor, which utilizes a K-nearest-neighbor algorithm. The similarity metric between any two sequences is 1- E/L, where L is a parameter, preferably 8-10 and E is the edit distance between the two sequences, taking into account only the first L nucleotides of each sequence. If the K-nearest-neighbor scores of two or more locations on the GAM FOLDED PRECURSOR RNA (Fig. 1) are not significantly different, these locations are further ranked by a Bayesian model, similar to the one described hereinabove.

[0222] In accordance with an embodiment of the present invention, scores of two or more of the abovementioned classifiers or predictors are integrated, yielding an integrated score for each potential GAM RNA. As an example, Fig. 6C illustrates an integration of scores from two classifiers, a 3' end recognition classifier and a 5' end recognition classifier, the scores of which are integrated to yield an integrated score. Most preferably, the INTEGRATED SCORE of

Fig. 6C preferably implements a "best-of-breed" approach employing a pair of classifiers and accepting only "potential GAM RNAs" that score highly on one of the abovementioned "EDIT DISTANCE" or "TWO PHASE" predictors. In this context, "high scores" means scores that have been demonstrated to have low false positive value when scoring known miRNA oligonucleotides. Alternatively, the INTEGRATED SCORE may be derived from operation of more or less than two classifiers.

[0223] The INTEGRATED SCORE is evaluated as follows: (a) the "potential GAM RNA" having the highest score is preferably taken to be the most probable GAM RNA, and (b) if the integrated score of this most probable GAM RNA is higher than a pre-defined threshold, then the most probable GAM RNA is accepted as a PREDICTED GAM RNA. Preferably, this evaluation technique is not limited to the highest scoring potential GAM RNA.

[0224] In a preferred embodiment of the present invention, PREDICTED GAM RNAs comprising a low complexity nucleotide sequence (e.g., ATATATA) may optionally be filtered out, because there is a high probability that they are part of a repeated element in the DNA, and are therefore not functional, as is known in the art. For each PREDICTED

GAM RNA sequence, the number of occurrences of each two nt combination (AA, AT, AC) comprised in that sequence is counted. PREDICTED GAM RNA sequences where the sum of the two most probable combinations is higher than a threshold, preferably 8-10, are filtered out. As an example, when the threshold is set such that 2% of the known miRNA oligonucleotides are filtered out, 30% of the predicted GAM RNAs are filtered out.

[0225] Reference is now made to Fig. 7A, which is a simplified block diagram of a preferred implementation of the target gene binding site detector 118 described hereinabove with reference to Fig. 2. The goal of the target gene binding site detector 118 is to detect one or more binding sites located in 3'UTRs of the mRNA of a known gene, such as BINDING SITE I, BINDING SITE II and BINDING SITE III (Fig. 1), the nucleotide sequence of which binding sites is partially or fully complementary to a GAM RNA, thereby determining that the abovementioned known gene is a target gene of the GAM RNA.

[0226] The target gene binding site detector 118 (Fig. 2) receives a plurality of Dicer-cut sequences from hairpin structures 140 (Fig. 6A) and a plurality of potential target gene sequences 142, which are derived from sequenced DNA data

104 (Fig. 2).

[0227] The target gene binding site detector training & validation functionality 128 (Fig. 3) is operative to train the target gene binding site detector 118 on known miRNA oligonucleotides and their respective target genes and to build a background model for an evaluation of the probability of achieving similar results randomly (P value) for the target gene binding site detector 118 results. The target gene binding site detector training & validation functionality 128 constructs the model by analyzing both heuristically and computationally the results of the target gene binding site detector 118.

[0228] Following operation of target gene binding site detector training & validation functionality 128 (Fig. 3), the target gene binding site detector 118 is operative to detect a plurality of potential novel target genes having binding site/s 144, the nucleotide sequence of which is partially or fully complementary to that of each of the plurality of Dicer-cut sequences from hairpin structures 140. Preferred operation of the target gene binding site detector 118 is further described hereinbelow with reference to Fig. 7B.

[0229] Reference is now made to Fig. 7B, which is a simplified

flowchart illustrating a preferred operation of the target gene binding site detector 118 of Fig. 2.

[0230] In an embodiment of the present invention, the target gene binding site detector 118 first compares nucleotide sequences of each of the plurality of Dicer-cut sequences from hairpin structures 140 (Fig. 6A) to the potential target gene sequences 142 (Fig. 7A), such as 3' side UTRs of known mRNAs, in order to find crude potential matches. This step may be performed using a simple alignment algorithm such as BLAST.

[0231] Then, the target gene binding site detector 118 filters these crude potential matches, to find closer matches, which more closely resemble published miRNA oligonucleotide binding sites.

[0232] Next, the target gene binding site detector 118 expands the nucleotide sequences of the 3'UTR binding site found by the sequence comparison algorithm (e.g. BLAST or EDIT DISTANCE). A determination is made whether any subsequence of the expanded sequence may improve the match. The best match is considered the alignment.

[0233] Free-energy and spatial structure are computed for the resulting binding sites. Calculation of spatial structure may be performed by a secondary structure folding algo-

rithm based on free-energy minimization, such as the MFOLD algorithm described in Mathews et al. (J. Mol. Biol. 288: 911-940 (1999)) and Zuker (Nucleic Acids Res. 31: 3406-3415 (2003)), the disclosure of which is hereby incorporated by reference. Free-energy, spatial structure and the above preferences are reflected in scoring. The resulting scores are compared with scores characterstic of known binding sites of published miRNA oligonucleotides, and each binding site is given a score that reflects its resemblance to these known binding sites.

[0234] Finally, the target gene binding site detector 118 analyzes the spatial structure of the binding site. Each 3'UTR-GAM oligonucleotide pair is given a score. Multiple binding sites of the same GAM oligonucleotides to a 3'UTR are given higher scores than those that bind only once to a 3'UTR.

[0235] In a preferred embodiment of the present invention, performance of the target gene binding site detector 118 may be improved by integrating several of the abovementioned logical steps, using the methodology described hereinbelow.

[0236] For each of the Dicer-cut sequence from hairpin structures 140, its starting segment, e.g. a segment compris-

ing the first 8 nts from its 5' end, is obtained. For each starting segment, all of the 9 nt segments that are highly complementary to the starting segment are calculated. These calculated segments are referred to here as "potential binding site end segments". In a preferred embodiment of the present invention, for each 8 nt starting segment, the potential binding site end segments are all 9 nt segments whose complementary sequence contains a 7-9 nt sub-sequence that is not different from the starting segment by more than an insertion, deletion or replacement of one nt. Calculation of potential binding site end segments is preferably performed by a pre-processing tool that maps all possible 8 nt segments to their respective 9 nt segments.

[0237] Next, the mRNAs 3'UTRs is parsed into all the segments, with the same length as the potential binding site end segments, preferably 9 nt segments, comprised in the 3'UTR. Location of each such segment is noted, stored in a performance-efficient data structure and compared to the potential binding site end segments calculated in the previous step.

[0238] The target gene binding site detector 118 then expands the binding site sequence, preferably in the binding site 5'

direction (i.e. immediately upstream), assessing the degree of its alignment to the Dicer-cut sequence from hairpin structures 140. Preferably, an alignment algorithm is implemented which uses specific weighting parameters based on an analysis of known miRNA oligonucleotide binding sites. As an example, it is apparent that a good match of the 3' end of the binding site is critically important, a match of the 5' end is less important but can compensate for a small number of mismatches at the 3' end of the binding site, and a match of the middle portion of the binding site is much less important.

[0239] Next, the number of binding sites found in a specific 3'UTR, the degree of alignment of each of these binding sites, and their proximity to each other are assessed and compared to these properties found in known binding sites of published miRNA oligonucleotides. In a preferred embodiment, the fact that many of the known binding sites are clustered is used to evaluate the P value of obtaining a cluster of a few binding sites on the same target gene 3'UTR in the following way. It scans different score thresholds and calculates for each threshold the number and positions of possible binding sites with a score above the threshold. It then gets a P value for each threshold

from a preprocessed calculated background matrix, described hereinbelow, and a number and positions of binding sites combination. The output score for each Dicer-cut sequences from hairpin structures 140 and potential target gene sequences 142 is the minimal P value, normalized with the number of threshold trails using a Bernoulli distribution. A preference of low P value pairs is made.

[0240] As mentioned hereinabove, for each target gene, a preprocessed calculated background matrix is built. The matrix includes rows for each number of miRNA oligonucleotide binding sites (in the preferred embodiment, the matrix includes 7 rows to accommodate 0 to 6 binding sites), and columns for each different score threshold (in the preferred embodiment, the matrix includes 5 columns for 5 different thresholds). Each matrix cell, corresponding to a specific number of binding sites and thresholds, is set to be the probability of getting equal or higher number binding sites and an equal or higher score using random 22 nt-long sequences with the same nucleotide distribution as known miRNA oligonucleotides (29.5% T, 24.5% A, 25% G and 21% C). Those probabilities are calculated by running the above procedure for 10000 random

sequences that preserved the known miRNA nucleotide distribution (these sequence will be also referred to as miRNA oligonucleotide random sequences). The P value can be estimated as the number of random sequences that obeys the matrix cell requirement divided by the total number of random sequences (10000). In the preferred embodiment, 2 matrices are calculated. The P values of the second matrix are calculated under a constraint that at least two of the binding site positions are under a heuristically-determined constant value. The values of the second matrix are calculated without this constraint. The target gene binding site detector 118 uses the second matrix if the binding site positions agree with the constraint. Otherwise, it uses the first. In an alternative embodiment, only one matrix is calculated without any constraint on the binding sites positions.

[0241] A test performed using the target gene binding site detector 118 shows that all of the known miRNA oligonucleotide target genes are found using this algorithm with a P value of less than 0.5%. Running known miRNA oligonucleotides against 3400 potential 3'UTR of target gene sequences yields on average 32 target genes for each miRNA oligonucleotide with a P value less than 0.5%,

while background sequences, as well as inverse or complement sequence of known miRNA oligonucleotide (which preserve their high order sequence statistics) found, as expected, 17 target genes on average. This result reflects that the algorithm has the ability to detect real target genes with 47% accuracy.

[0242] Finally, orthology data may optionally be used to further prefer binding sites based on their conservation. Preferably, this may be used in cases such as (a) where both the target mRNA and miRNA oligonucleotide have orthologues in another organism, e.g. Human-Mouse orthology, or (b) where a miRNA oligonucleotide (e.g. viral miRNA oligonucleotide) targets two mRNAs in orthologous organisms. In such cases, binding sites that are conserved are preferred.

[0243] In accordance with another preferred embodiment of the present invention, binding sites may be searched by a reverse process. Sequences of K (preferably 22) nucleotides in a UTR of a target gene are assessed as potential binding sites. A sequence comparison algorithm, such as BLAST or EDIT DISTANCE variant, is then used to search elsewhere in the genome for partially or fully complementary sequences that are found in known miRNA oligonucleotides or computationally-predicted GAM oligonu-

cleotides. Only complementary sequences that meet predetermined spatial structure and free-energy criteria as described hereinabove, are accepted. Clustered binding sites are strongly preferred and potential binding sites and potential GAM oligonucleotides that occur in evolutionarily-conserved genomic sequences are also preferred. Scoring of candidate binding sites takes into account free-energy and spatial structure of the binding site complexes, as well as the aforesaid preferences.

[0244] The 3'UTR of each bacterial gene is extracted from the 500 nts that lay downstream to the gene-coding region. Care is taken that the extracted 3'UTR is not partly covered by the predicted 5'UTR of the next gene-coding region, considered 300 nts upstream. This method is applied on known (not hypothetical) bacterial genes of completed pathogenic eubacterial genomes taken from the updated NCBI Ref_seq database on 17 Mar 2004.

[0245] Reference is now made to Fig. 8, which is a simplified flowchart illustrating a preferred operation of the function & utility analyzer 120 described hereinabove with reference to Fig. 2. The goal of the function & utility analyzer 120 is to determine if a potential target gene is in fact a valid clinically useful target gene. Since a potential novel

GAM oligonucleotide binding a binding site in the UTR of a target gene is understood to inhibit expression of that target gene, and if that target gene is shown to have a valid clinical utility, then in such a case it follows that the potential novel oligonucleotide itself also has a valid useful function which is the opposite of that of the target gene.

[0246] The function & utility analyzer 120 preferably receives as input a plurality of potential novel target genes having binding site/s 144 (Fig. 7A), generated by the target gene binding site detector 118 (Fig. 2). Each potential oligonucleotide is evaluated as follows: First, the system checks to see if the function of the potential target gene is scientifically well established. Preferably, this can be achieved bioinformatically by searching various published data sources presenting information on known function of proteins. Many such data sources exist and are published, as is well known in the art. Next, for those target genes the function of which is scientifically known and is well documented, the system then checks if scientific research data exists which links them to known diseases. For example, a preferred embodiment of the present invention utilizes the OMIM(TM) (Hamosh et al, 2002) database published by

NCBI, which summarizes research publications relating to genes which have been shown to be associated with diseases. Finally, the specific possible utility of the target gene is evaluated. While this process too may be facilitated by bioinformatic means, it might require manual evaluation of published scientific research regarding the target gene, in order to determine the utility of the target gene to the diagnosis and or treatment of specific disease. Only potential novel oligonucleotides, the target genes of which have passed all three examinations, are accepted as novel oligonucleotide.

[0247] Reference is now made to Fig. 9, which is a simplified diagram describing each of a plurality of novel bioinformatically-detected regulatory polynucleotide referred to in this Table as the Genomic Record (GR) polynucleotide. GR encodes an operon-like cluster of novel miRNA-like oligonucleotides, each of which in turn modulates expression of at least one target gene. The function and utility of at least one target gene is known in the art.

[0248] The GR PRECURSOR is a novel, bioinformatically-detected, regulatory, non-protein-coding polynucleotide. The method by which the GR PRECURSOR is detected is described hereinabove with additional reference to Figs. 1-9.

[0249] GR PRECURSOR is preferably encoded by the bacterial genome and contains a cluster of novel bacterial oligonucleotides, which preferably bind to human target genes or to bacterium genes. Alternatively or additionally, GR PRECURSOR is encoded by the human genome and contains a cluster of novel human oligonucleotides, which preferably bind to bacterial target genes or to human genes.

[0250] The GR PRECURSOR encodes GR PRECURSOR RNA that is typically several hundred to several thousand nts long. The GR PRECURSOR RNA folds spatially, forming the GR FOLDED PRECURSOR RNA. It is appreciated that the GR FOLDED PRECURSOR RNA comprises a plurality of what is known in the art as hairpin structures. Hairpin structures result from the presence of segments of the nucleotide sequence of GR PRECURSOR RNA in which the first half of each such segment has a nucleotide sequence which is at least a partial, and sometimes an accurate, reverse-complement sequence of the second half thereof, as is well known in the art.

[0251] The GR FOLDED PRECURSOR RNA is naturally processed by cellular enzymatic activity into a plurality of separate GAM precursor RNAs herein schematically represented by GAM1 FOLDED PRECURSOR RNA through GAM3 FOLDED

PRECURSOR RNA. Each GAM folded precursor RNA is a hairpin-shaped RNA segment, corresponding to GAM FOLDED PRECURSOR RNA of Fig. 1.

[0252] The abovementioned GAM folded precursor RNAs are diced by DICER COMPLEX of Fig. 1, yielding short RNA segments of about 22 nts in length schematically represented by GAM1 RNA through GAM3 RNA. Each GAM RNA corresponds to GAM RNA of Fig. 1. GAM1 RNA, GAM2 RNA and GAM3 RNA each bind complementarily to binding sites located in the untranslated regions of their respective target genes, designated GAM1 TARGET RNA, GAM2 TARGET RNA and GAM3 TARGET RNA, respectively. These target binding sites correspond to BINDING SITE I, BINDING SITE II and BINDING SITE III of Fig. 1. The binding of each GAM RNA to its target RNA inhibits the translation of its respective target proteins, designated GAM1 TARGET PROTEIN, GAM2 TARGET PROTEIN and GAM3 TARGET PROTEIN, respectively.

[0253] It is appreciated that the specific functions, and accordingly the utilities, of the GR polynucleotide are correlated with and may be deduced from the identity of the target genes that are inhibited by GAM RNAs that are present in the operon-like cluster of the polynucleotide. Thus, for

the GR polynucleotide, schematically represented by GAM1 TARGET PROTEIN through GAM3 TARGET PROTEIN that are inhibited by the GAM RNA. The function of these target genes is elaborated in Table 8, hereby incorporated herein.

[0254] Reference is now made to Fig. 10, which is a block diagram illustrating different utilities of oligonucleotide of the novel group of oligonucleotides of the present invention referred to here as GAM oligonucleotides and GR polynucleotides. The present invention discloses a first plurality of novel oligonucleotides referred to here as GAM oligonucleotides and a second plurality of operon-like polynucleotides referred to here as GR polynucleotides, each of the GR polynucleotide encoding a plurality of GAM oligonucleotides. The present invention further discloses a very large number of known target genes, which are bound by, and the expression of which is modulated by each of the novel oligonucleotides of the present invention. Published scientific data referenced by the present invention provides specific, substantial, and credible evidence that the abovementioned target genes modulated by novel oligonucleotides of the present invention, are associated with various diseases. Specific novel oligonu-

cleotides of the present invention, target genes thereof and diseases associated therewith, are described herein-below with reference to Tables 1 through 12. It is therefore appreciated that a function of GAM oligonucleotides and GR polynucleotides of the present invention is modulation of expression of target genes related to known bacterial diseases, and that therefore utilities of novel oligonucleotides of the present invention include diagnosis and treatment of the abovementioned diseases.

[0255] Fig. 10 describes various types of diagnostic and therapeutic utilities of novel oligonucleotides of the present invention. A utility of novel oligonucleotide of the present invention is detection of GAM oligonucleotides and of GR polynucleotides. It is appreciated that since GAM oligonucleotides and GR polynucleotides modulate expression of disease related target genes, that detection of expression of GAM oligonucleotides in clinical scenarios associated with said bacterial diseases is a specific, substantial and credible utility. Diagnosis of novel oligonucleotides of the present invention may preferably be implemented by RNA expression detection techniques, including but not limited to biochips, as is well known in the art. Diagnosis of expression of oligonucleotides of the present invention may

be useful for research purposes, in order to further understand the connection between the novel oligonucleotides of the present invention and the abovementioned related bacterial diseases, for disease diagnosis and prevention purposes, and for monitoring disease progress.

[0256] Another utility of novel oligonucleotides of the present invention is anti-GAM therapy, a mode of therapy which allows up regulation of a bacterial disease-related target gene of a novel GAM oligonucleotide of the present invention, by lowering levels of the novel GAM oligonucleotide which naturally inhibits expression of that target gene. This mode of therapy is particularly useful with respect to target genes which have been shown to be under-expressed in association with a specific bacterial disease. Anti-GAM therapy is further discussed hereinbelow with reference to Figs. 11A and 11B.

[0257] A further utility of novel oligonucleotides of the present invention is GAM replacement therapy, a mode of therapy which achieves down regulation of a bacterial disease related target gene of a novel GAM oligonucleotide of the present invention, by raising levels of the GAM which naturally inhibits expression of that target gene. This mode

of therapy is particularly useful with respect to target genes which have been shown to be over-expressed in association with a specific bacterial disease. GAM replacement therapy involves introduction of supplementary GAM products into a cell, or stimulation of a cell to produce excess GAM products. GAM replacement therapy may preferably be achieved by transfecting cells with an artificial DNA molecule encoding a GAM which causes the cells to produce the GAM product, as is well known in the art.

[0258] Yet a further utility of novel oligonucleotides of the present invention is modified GAM therapy. Disease conditions are likely to exist, in which a mutation in a binding site of a GAM RNA prevents natural GAM RNA to effectively bind inhibit a bacterial disease related target gene, causing up regulation of that target gene, and thereby contributing to the disease pathology. In such conditions, a modified GAM oligonucleotides is designed which effectively binds the mutated GAM binding site, i.e. is an effective anti-sense of the mutated GAM binding site, and is introduced in disease effected cells. Modified GAM therapy is preferably achieved by transfecting cells with an artificial DNA molecule encoding the modified GAM which causes the cells to produce the modified GAM product, as

is well known in the art.

[0259] Reference is now made to Figs. 11A and 11B, which are simplified diagrams which when taken together illustrate anti-GAM therapy mentioned hereinabove with reference to Fig. 10. A utility of novel GAMs of the present invention is anti-GAM therapy, a mode of therapy which allows up regulation of a bacterial disease-related target gene of a novel GAM of the present invention, by lowering levels of the novel GAM which naturally inhibits expression of that target gene. Fig. 11A shows a normal GAM inhibiting translation of a target gene by binding of GAM RNA to a BINDING SITE found in an untranslated region of GAM TARGET RNA, as described hereinabove with reference to Fig. 1.

[0260] Fig. 11B shows an example of anti-GAM therapy. ANTI-GAM RNA is short artificial RNA molecule the sequence of which is an anti-sense of GAM RNA. Anti-GAM treatment comprises transfecting diseased cells with ANTI-GAM RNA, or with a DNA encoding thereof. The ANTI-GAM RNA binds the natural GAM RNA, thereby preventing binding of natural GAM RNA to its BINDING SITE. This prevents natural translation inhibition of GAM TARGET RNA by GAM RNA, thereby up regulating expression of GAM TARGET

PROTEIN.

[0261] It is appreciated that anti-GAM therapy is particularly useful with respect to target genes which have been shown to be under-expressed in association with a specific bacterial disease.

[0262] Furthermore, anti-GAM therapy is particularly useful, since it may be used in situations in which technologies known in the art as RNAi and siRNA can not be utilized. As in known in the art, RNAi and siRNA are technologies which offer means for artificially inhibiting expression of a target protein, by artificially designed short RNA segments which bind complementarily to mRNA of said target protein. However, RNAi and siRNA can not be used to directly up regulate translation of target proteins.

[0263] Reference is now made to Fig. 12A, which is a bar graph illustrating performance results of the hairpin detector 114 (Fig. 2) constructed and operative in accordance with a preferred embodiment of the present invention.

[0264] Fig. 12A illustrates efficacy of several features used by the hairpin detector 114 to detect GAM FOLDED PRECURSOR RNAs (Fig. 1). The values of each of these features is compared between a set of published miRNA precursor oligonucleotides, represented by shaded bars, and a set of

random hairpins folded from the human genome denoted hereinbelow as a hairpin background set, represented by white bars. The published miRNA precursor oligonucleotides set is taken from RFAM database, Release 2.1 and includes 148 miRNA oligonucleotides from H.Sapiens. The background set comprises a set of 10,000 hairpins folded from the human genome.

[0265] It is appreciated that the hairpin background set is expected to comprise some valid, previously undetected hairpin-shaped miRNA precursor-like GAM FOLDED PRECURSOR RNAs of the present invention, and many hairpin-shaped sequences that are not hairpin-shaped miRNA-like precursors.

[0266] For each feature, the bars depict the percent of known miRNA hairpin precursors (shaded bars) and the percent of background hairpins (white bars) that pass the threshold for that feature. The percent of known miRNA oligonucleotides that pass the threshold indicates the sensitivity of the feature, while the corresponding background percent implies the specificity of the feature, although not precisely, because the background set comprises both true and false examples.

[0267] The first bar pair, labeled Thermodynamic Stability Selec-

tion, depicts hairpins that have passed the selection of "families" of closely related hairpin structures, as described hereinabove with reference to Fig. 5B.

[0268] The second bar pair, labeled Hairpin Score, depicts hairpins that have been selected by hairpin detector 114 (Fig. 5B), regardless of the "families" selection.

[0269] The third bar pair, labeled Conserved, depicts hairpins that are conserved in human, mouse and rat, (UCSC Goldenpath (TM) HG16 database).

[0270] The fourth bar pair, labeled Expressed, depicts hairpins that are found in EST blocks.

[0271] The fifth bar pair, labeled Integrated Selection, depicts hairpin structures predicted by a preferred embodiment of the present invention to be valid GAM PRECURSORs. In a preferred embodiment of the present invention, a hairpin may be considered to be a GAM PRECURSOR if its hairpin detector score is above 0, and it is in one of the following groups: a) in an intron and conserved or b) in an intergenic region and conserved or c) in an intergenic region and expressed, as described below. Further filtering of GAM precursor may be obtained by selecting hairpins with a high score of Dicer-cut location detector 116 as described hereinabove with reference to Figs. 6A-6C, and

with predicted miRNA oligonucleotides, which pass the low complexity filter as described hereinabove, and whose targets are selected by the target gene binding site detector 118 as described hereinabove with reference to Figs. 7A-7B.

[0272] It is appreciated that these results validate the sensitivity and specificity of the hairpin detector 114 (Fig. 2) in identifying novel GAM FOLDED PRECURSOR RNAs, and in effectively distinguishing them from the abundant hairpins found in the genome.

[0273] Reference is now made to Fig. 12B, which is a line graph illustrating accuracy of a Dicer-cut location detector 116 (Fig. 2) constructed and operative in accordance with a preferred embodiment of the present invention.

[0274] To determine the accuracy of the Dicer-cut location detector 116, a stringent training and test set was chosen from the abovementioned set of 440 known miRNA oligonucleotides, such that no two miRNA oligonucleotides in the set are homologous. This was performed to get a lower bound on the accuracy and avoid effects of similar known miRNA oligonucleotides appearing in both the training and test sets. On this stringent set of size 204, mfold cross validation with k=3 was performed to

determine the percent of known miRNA oligonucleotides in which the Dicer-cut location detector 116 described hereinabove predicted the correct miRNA oligonucleotide up to two nucleotides from the correct location. The accuracy of the TWO PHASED predictor is depicted in the graph. The accuracy of the first phase of the TWO PHASED predictor is depicted by the upper line, and that of both phases of the TWO PHASED predictor is depicted by the lower line. Both are binned by the predictor score, where the score is the score of the first stage.

[0275] It is appreciated that these results validate the accuracy of the Dicer-cut location detector 116.

[0276] Reference is now made to Fig. 12C, which is a bar graph illustrating the performance results of the target gene binding site detector 118 (Fig. 7A) constructed and operative in accordance with a preferred embodiment of the present invention.

[0277] Fig. 12C illustrates specificity and sensitivity of the target gene binding site detector 118. The values presented are the result of testing 10000 artificial miRNA oligonucleotide sequences (random 22 nt sequences with the same base composition as published miRNA oligonucleotide sequence). Adjusting the threshold parameters to

fulfill 90% sensitivity of validated, published miRNA-3'UTR pairs, requires the P VAL of potential target gene sequences-Dicer-cut sequences to be less than 0.01 and also the P VAL of potential target ortholog gene sequences-Dicer-cut sequences to be less than 0.05. The target gene binding site detector 118 can filter out 99.7% of potential miRNA/gene pairs, leaving only the 0.3% that contain the most promising potential miRNA/gene pairs. Limiting the condition for the P VAL of potential target ortholog gene sequences-Dicer-cut sequences to be less than 0.01 reduces the sensitivity ratio to 70% but filters out more then 50% of the remaining 0.3%, to a final ratio of less than 0.15%.

[0278] It is appreciated that these results validate the sensitivity and specificity of the target gene binding site detector 118.

[0279] Reference is now made to Fig. 13, which is a summary table of laboratory results validating the expression of 29 novel human GAM RNA oligonucleotides in HeLa cells or, alternatively, in liver or thymus tissues detected by the bioinformatic oligonucleotide detection engine 100 (Fig. 2).

[0280] As a positive control, we used a reference set of eight

known human miRNA oligonucleotides: hsa-MIR-21; hsa-MIR-27b; hsa-MIR-186; hsa-MIR-93; hsa-MIR-26a; hsa-MIR-191; hsa-MIR-31; and hsa-MIR-92. All positive controls were successfully validated by sequencing.

[0281] The table of Fig. 13 lists all GAM RNA predictions whose expression was validated. The field "Primer Sequence" contains the "specific" part of the primer; the field "Sequenced sequence" represents the nucleotide sequence detected by cloning (excluding the hemispecific primer sequence); the field "Predicted GAM RNA" contains the GAM RNA predicted sequence; the field "Distance indicate the distance from Primer; the number of mismatches between the "specific" region of the primer and the corresponding part of the GAM RNA sequence; the field "GAM Name" contains GAM RNA PRECURSOR ID followed by "A" or "B", which represents the GAM RNA position on the precursor as elaborated in the attached Tables.

[0282] A primer was designed such that its first half, the 5' region, is complementary to the adaptor sequence and its second half, the 3' region, anneals to the 5' terminus of GAM RNA sequence, yielding a hemispecific primer (as elaborated hereinbelow in the Methods section). A sample of 13 predicted GAM RNA sequences was examined by

PCR using hemispecific primers and a primer specific to the 3' adaptor. PCR products were cloned into plasmid vectors and then sequenced. For all 13 predicted GAM RNA sequences, the GAM RNA sequence found in the hemispecific primer plus the sequence observed between the hemispecific primer and the 3' adaptor was completely included in the expected GAM RNA sequence (rows 1-7, and 29). The rest are GAM RNA predictions that were verified by cloning and sequencing, yet, by using a primer that was originally designed for a slightly different prediction.

[0283] It is appreciated that failure to detect a predicted oligonucleotide in the lab does not necessarily indicate a mistaken bioinformatic prediction. Rather, it may be due to technical sensitivity limitation of the lab test, or because the predicted oligonucleotides are not expressed in the tissue examined, or at the development phase tested. The observed GAM RNAs may be strongly expressed in HeLa cells while the original GAM RNAs are expressed at low levels in HeLa cells or not expressed at all. Under such circumstances, primer sequences containing up to three mismatches from a specific GAM RNA sequence may amplify it. Thus, we also considered cases in which differ-

ences of up to 3 mismatches in the hemispecific primer occur.

[0284] The 3' terminus of observed GAM RNA sequences is often truncated or extended by one or two nucleotides. Cloned sequences that were sequenced from both 5' and 3' termini have an asterick appended to the row number.

[0285] Interestingly, the primer sequence followed by the observed cloned sequence is contained within five GAM RNA sequences of different lengths, and belong to 24 precursors derived from distinct loci (Row 29). Out of these, one precursor appears four times in the genome and its corresponding GAM Names are 351973-A, 352169-A, 352445-A and 358164-A.

[0286] The sequence presented in Row 29 is a representative of the group of five GAM RNAs. The full list of GAM RNA sequences and their corresponding precursors is as follows (each GAM RNA sequence is followed by the GAM Name): TCACTGCAACCTCCACCTCCCA (352092, 352651,355761),TCACTGCAACCTCCACCTCCCG (351868, 352440, 351973, 352169, 352445, 358164, 353737, 352382, 352235, 352232, 352268, 351919, 352473, 352444, 353638, 353004, 352925, 352943), TCACTGCAACCTCCACCTCCTG

(358311),TCACTGCAACCTCCACCTTCAG (353323), and TCACTGCAACCTCCACCTTCCG (353856).

[0287] METHOD SECTION

[0288] CELL LINES

[0289] Three common human cell lines, obtained from Dr. Yonat Shemer at Soroka Medical Center, Be'er Sheva, Israel, were used for RNA extraction; Human Embryonic Kidney HEK-293 cells, Human Cervix Adenocarcinoma HeLa cells and Human Prostate Carcinoma PC3cells.

[0290] RNA PURIFICATION

[0291] Several sources of RNA were used to prepare libraries:

[0292] Total HeLa S100 RNA was prepared from HeLa S100 cellular fraction (4C Biotech, Belgium) through an SDS (1%)-Proteinase K (200g/ml) 30 minute incubation at 37 C followed by an acid Phenol-Chloroform purification and isopropanol precipitation (Sambrook et al; Molecular Cloning- A Laboratory Manual).

[0293] Total HeLa, HEK-293 and PC3 cell RNA was prepared using the standard Tri-Reagent protocol (Sigma) according to the manufacturer's instructions, except that 1 volume of isopropanol was substituted with 3 volumes of ethanol.

[0294] Nuclear and Cytoplasmic RNA was prepared from HeLa or

HEK-293 cells in the following manner:

[0295] Cell were washed and harvested in ice-cold PBS and precipitated in a swing-out rotor at 1200 rpm at 4 C for 5 minutes. Pellets were loosened by gentle vortexing. 4ml of "NP40 lysis buffer" (10mM TrisHCl, 5mM MgCl2, 10mM NaCl, 0.5% Nonidet P40 , 1mM Spermidine, 1mM DTT, 140U/ml rRnasine) was then added per 5*107 cells. Cells and lysis buffer were incubated for 5 minutes on ice and centrifuged in a swing-out rotor at 500xg at 4 C for 5 minutes. Supernatant, termed cytoplasm, is carefully removed to a tube containing SDS (1% final) and proteinase-K (200 g/ml final). Pellet, termed nuclear fraction, is rewashed and incubated with a similar amount of fresh lysis buffer. Lysis is monitored visually under a microscope at this stage, typically for 5 minutes. Nuclei are pelleted in a swing-out rotor at 500xg at 4 C for 5 minutes. Supernatant is pooled, incubated at 37 C for 30 minutes, Phenol/Chloroform-extracted, and RNA is alcohol-precipitated (Sambrook et al). Nuclei are loosened and then homogenized immediately in >10 volumes of Tri-Reagent (Sigma). Nuclear RNA is then prepared according to the manufacturer's instructions.

[0296] TOTAL TISSUE RNA

[0297] Total tissue RNA was obtained from Ambion USA, and included Human Liver, Thymus, Placenta, Testes and Brain.

[0298] RNA SIZE FRACTIONATION

[0299] RNA used for libraries was always size-fractionated. Fractionation was done by loading up to 500 microgram RNA per YM100 Amicon Microcon column (Millipore) followed by a 500xg centrifugation for 40 minutes at 4 C. Flow-through "YM100" RNA is about one quarter of the total RNA and was used for library preparation or fractionated further by loading onto a YM30 Amicon Microcon column (Millipore) followed by a 13,500xg centrifugation for 25 minutes at 4 C. Flow-through "YM30" was used for library preparation "as is" and consists of less than 0.5% of total RNA. Additional size fractionation was achieved during library preparation.

[0300] LIBRARY PREPARATION

[0301] Two types of cDNA libraries, designated "One-tailed" and "Ligation", were prepared from the one of the abovementioned fractionated RNA samples. RNA was dephosphorylated and ligated to an RNA (designated with lowercase letters)-DNA (designated with UPPERCASE letters) hybrid 5'-phosphorylated, 3' idT blocked 3'-adapter

(5'-P-uuuAACCGCATCCTTCTC-idT-3' Dharmacon # P-002045-01-05) (as elaborated in Elbashir et al., Genes Dev. 15:188-200 (2001)) resulting in ligation only of RNase III type cleavage products. 3'-Ligated RNA was excised and purified from a half 6%, half 13% polyacrylamide gel to remove excess adapter with a Nanosep 0.2 microM centrifugal device (Pall) according to instructions, and precipitated with glycogen and 3 volumes of ethanol. Pellet was resuspended in a minimal volume of water.

[0302] For the "Ligation" library, a DNA (UPPERCASE)-RNA (lowercase) hybrid 5'-adapter (5'-TACTAATACGACTCACTaaa-3' Dharmacon # P-002046-01-05) was ligated to the 3'-adapted RNA, reverse transcribed with "EcoRI-RT": (5'-GACTAGCTGGAATTCAAGGATGCGGTTAAA-3'), PCR-amplified with two external primers essentially as in Elbashir et al. (2001), except that primers were "EcoRI-RT" and "PstI Fwd"(5'-CAGCCAACGCTGCAGATACGACTCACTAAA-3'). This PCR product was used as a template for a second round of PCR with one hemispecific and one external primer or with two hemispecific primers.

[0303] For the "One-tailed" library, the 3'-adapted RNA was an-

nealed to 20pmol primer "EcoRI RT" by heating to 70 C and cooling 0.1 C/sec to 30 C and then reverse-transcribed with Superscript II RT (according to manufacturer's instructions, Invitrogen) in a 20 microliters volume for 10 alternating 5 minute cycles of 37 C and 45 C. Subsequently, RNA was digested with 1 microliter 2M NaOH and 2mM EDTA at 65 C for 10 minutes. cDNA was loaded on a polyacrylamide gel, excised and gel-purified from excess primer as above (invisible, judged by primer run alongside) and resuspended in 13 microliters of water. Purified cDNA was then oligo-dC tailed with 400U of recombinant terminal transferase (Roche Molecular Biochemicals), 1 microliter 100 microM dCTP, 1 microliter 15mM CoCl2, and 4 microliters reaction buffer, to a final volume of 20 microliters for 15 minutes at 37 C. Reaction was stopped with 2 microliters 0.2M EDTA and 15 microliters 3M NaOAc pH 5.2. Volume was adjusted to 150 microliters with water, Phenol: Bromochloropropane 10:1 extracted and subsequently precipitated with glycogen and 3 volumes of ethanol. C-tailed cDNA was used as a template for PCR with the external primers "T3-PstBsg(G/I)18"(5'-AATTAACCCTCACTAAAGGCTGCAG GTGCAGGIGGGIIGGGIIGGGIIGN-3' where I stands for Ino-

sine and N for any of the 4 possible deoxynucleotides), and with "EcoRI Nested"(5'-GGAATTCAAGGATGCGGTTA-3'). This PCR product was used as a template for a second round of PCR with one hemispecific and one external primer or with two hemispecific primers.

[0304] PRIMER DESIGN AND PCR

[0305] Hemispecific primers were constructed for each predicted GAM RNA oligonucleotide by an in-house program designed to choose about half of the 5' or 3' sequence of the GAM RNA corresponding to a TM of about 30 -34 C constrained by an optimized 3' clamp, appended to the cloning adapter sequence (for "One-tailed" libraries, 5'-GGNNGGGNNG on the 5' end or TTTAACCGCATC-3' on the 3' end of the GAM RNA; for "Ligation" libraries, the same 3' adapter and 5'-CGACTCACTAAA on the 5' end of the GAM RNA). Consequently, a fully complementary primer of a TM higher than 60 C was created covering only one half of the GAM RNA sequence permitting the unbiased elucidation by sequencing of the other half.

[0306] For each primer, the following criteria were used: Primers were graded according to the TM of the primer half and the nucleotide content of 3 nucleotides of the 3' clamp

from worst to best, roughly: GGG-3' <CCC-3' <TTT-3'/AAA-3' <GG-3' <CC-3' <a TM lower than 30 < a TM higher than 34 <TT-3'/AA-3' <3G/C nucleotide combination <3 A/T nucleotide combination <any combination of two/three different nucleotides <any combination of three/three different nucleotides.

[0307] VALIDATION PCR PRODUCT BY SOUTHERN BLOT

[0308] GAM RNA oligonucleotides were validated by hybridization of Polymerase Chain Reaction (PCR)-product Southern blots with a probe to the predicted GAM RNA.

[0309] PCR product sequences were confirmed by Southern blot (Southern E.M., Biotechnology 1992,24:122-139 (1975)) and hybridization with DNA oligonucleotide probes synthesized as complementary (antisense) to predicted GAM RNA oligonucleotides. Gels were transferred onto a Biodyne PLUS 0.45m (Pall) positively charged nylon membrane and UV cross-linked. Hybridization was performed overnight with DIG-labeled probes at 42 C in DIG Easy-Hyb buffer (Roche). Membranes were washed twice with 2xSSC and 0.1% SDS for 10 minutes at 42 C and then washed twice with 0.5xSSC and 0.1% SDS for 5 min at 42 C. The membrane was then developed by using a DIG luminescent detection kit (Roche) using anti-DIG and CSPD

reaction, according to the manufacturer's protocol. All probes were prepared according to the manufacturer's (Roche Molecular Biochemicals) protocols: Digoxigenin (DIG) labeled antisense transcripts were prepared from purified PCR products using a DIG RNA labeling kit with T3 RNA polymerase. DIG-labeled PCR was prepared by using a DIG PCR labeling kit. 3'-DIG-tailed oligo ssDNA antisense probes, containing DIG-dUTP and dATP at an average tail length of 50 nts were prepared from 100pmole oligonucleotides with the DIG Oligonucleotide Labeling Kit. Control reactions contained all of the components of the test reaction except library template.

[0310] VALIDATION OF PCR PRODUCT BY NESTED PCR ON THE LIGATION

[0311] To further validate predicted GAM PCR product sequence derived from hemi-primers, a PCR-based diagnostic technique was devised to amplify only those products containing at least two additional nucleotides of the non hemi-primer defined part of the predicted GAM RNA oligonucleotide. In essence, a diagnostic primer was designed so that its 3' end, which is the specificity determining side, was identical to the desired GAM RNA oligonucleotide, 2-10 nts (typically 4-7, chosen for maximum specificity)

further into its 3' end than the nucleotide stretch primed by the hemi-primer. The hemi-primer PCR product was first ligated into a T-cloning vector (pTZ57/T or pGEM-T) as described hereinabove. The ligation reaction mixture was used as template for the diagnostic PCR under strict annealing conditions with the new diagnostic primer in conjunction with a general plasmid-homologous primer, resulting in a distinct ~200 base-pair product. This PCR product can be directly sequenced, permitting the elucidation of the remaining nucleotides up to the 3' of the mature GAM RNA oligonucleotide adjacent to the 3' adapter. Alternatively, following analysis of the diagnostic PCR reaction on an agarose gel, positive ligation reactions (containing a band of the expected size) were transformed into E. coli. Using this same diagnostic technique and as an alternative to screening by Southern blot colony hybridization, transformed bacterial colonies were screened by colony-PCR (Gussow, D. and Clackson, T, Nucleic Acids Res. 17:4000 (1989)) with the nested primer and the vector primer, prior to plasmid purification and sequencing.

[0312] VALIDATION OF PCR PRODUCT BY CLONING AND SEQUENCING

[0313] PCR products were inserted into pGEM-T (Promega) or

pTZ57/T (MBI Fermentas), heat-shock transformed into competent JM109 E. coli (Promega) and seeded on LB-Ampicilin plates with IPTG and Xgal. White and light blue colonies were transferred to duplicate gridded plates, one of which was blotted onto a membrane (Biodyne Plus, Pall) for hybridization with DIG tailed oligo probes (according to instructions, Roche) complementary to the expected GAM. Plasmid DNA from positive colonies was sequenced.

[0314] It is appreciated that the results summarize in Fig. 13 validate the efficacy of the bioinformatic oligonucleotide detection engine 100 of the present invention.

[0315] Reference is now made to Fig. 14A, which is a schematic representation of a novel human GR polynucleotide, located on chromosome 9, comprising 2 known human miRNA oligonucleotides - MIR24 and MIR23, and 2 novel GAM oligonucleotides, herein designated GAM7617 and GAM252 (later discovered by other researchers as hsa-mir-27b), all marked by solid black boxes. Fig. 14A also schematically illustrates 6 non-GAM hairpin sequences, and one non-hairpin sequence, all marked by white boxes, and serving as negative controls. By "non-GAM hairpin sequences" is meant sequences of a similar length to known miRNA precursor sequences, which form hairpin

secondary folding pattern similar to miRNA precursor hairpins, and yet which are assessed by the bioinformatic oligonucleotide detection engine 100 not to be valid GAM PRECURSOR hairpins. It is appreciated that Fig. 14A is a simplified schematic representation, reflecting only the order in which the segments of interest appear relative to one another, and not a proportional distance between the segments.

[0316] Reference is now made to Fig. 14B, which is a schematic representation of secondary folding of each of the MIRs and GAMs of the GR MIR24, MIR23, GAM7617 and GAM252, and of the negative control non-GAM hairpins, herein designated N2, N3, N252, N4, N6 and N7. N0 is a non-hairpin control, of a similar length to that of known miRNA precursor hairpins. It is appreciated that the negative controls are situated adjacent to and in between real miRNA oligonucleotides and GAM predicted oligonucleotides and demonstrates similar secondary folding patterns to that of known MIRs and GAMs.

[0317] Reference is now made to Fig. 14C, which is a picture of laboratory results of a PCR test upon a YM100 size-fractionated "ligation" library, utilizing a set of specific primer pairs located directly inside the boundaries of the

hairpins. Due to the nature of the library the only PCR amplifiable products can result from RNaseIII type enzyme cleaved RNA, as expected for legitimate hairpin precursors presumed to be produced by DROSHA (Lee et al, Nature 425 415-419, 2003). Fig. 14C demonstrates expression of hairpin precursors of known miRNA oligonucleotides hsa-mir23 and hsa-mir24, and of novel bioinformatically-detected GAM7617 and GAM252 hairpins predicted bioinformatically by a system constructed and operative in accordance with a preferred embodiment of the present invention. Fig. 14C also shows that none of the 7 controls (6 hairpins designated N2, N3, N23, N4, N6 and N7 and 1 non-hairpin sequence designated N0) were expressed. N252 is a negative control sequence partially overlapping GAM252.

[0318] In the picture, test lanes including template are designated "+" and the control lane is designated "-". The control reaction contained all the components of the test reaction except library template. It is appreciated that for each of the tested hairpins, a clear PCR band appears in the test ("+") lane, but not in the control ("-") lane.

[0319] Figs. 14A through 14C, when taken together validate the efficacy of the bioinformatic oligonucleotide detection en-

gine in: (a) detecting known miRNA oligonucleotides; (b) detecting novel GAM PRECURSOR hairpins which are found adjacent to these miRNA oligonucleotides, and which despite exhaustive prior biological efforts and bioinformatic detection efforts, went undetected; (c) discerning between GAM (or MIR) PRECURSOR hairpins, and non-GAM hairpins.

[0320] It is appreciated that the ability to discern GAM-hairpins from non-GAM-hairpins is very significant in detecting GAM oligonucleotides since hairpins are highly abundant in the genome. Other miRNA prediction programs have not been able to address this challenge successfully.

[0321] Reference is now made to Fig. 15A, which is an annotated sequence of an EST comprising a novel GAM oligonucleotides detected by the oligonucleotide detection system of the present invention. Fig. 15A shows the nucleotide sequence of a known human non-protein-coding EST (Expressed Sequence Tag), identified as EST72223. The EST72223 clone obtained from TIGR database (Kirkness and Kerlavage, 1997) was sequenced to yield the above 705bp transcript with a polyadenyl tail. It is appreciated that the sequence of this EST comprises sequences of one known miRNA oligonucleotide, identified as hsa-

MIR98, and of one novel GAM oligonucleotide referred to here as GAM25, detected by the bioinformatic oligonucleotide detection engine 100 (Fig. 2) of the present invention.

[0322] The sequences of the precursors of the known MIR98 and of the predicted GAM25 precursors are marked in bold, the sequences of the established miRNA 98 and of the predicted miRNA-like oligonucleotide GAM25 are underlined.

[0323] Reference is now made to Figs. 15B, 15C and 15D, which are pictures of laboratory results, which when taken together demonstrate laboratory confirmation of expression of the bioinformatically-detected novel oligonucleotide of Fig. 15A. In two parallel experiments, an enzymatically synthesized capped, EST72223 RNA transcript, was incubated with Hela S100 lysate for 0 minutes, 4 hours and 24 hours. RNA was subsequently harvested, run on a denaturing polyacrylamide gel, and reacted with either a 102 nt antisense MIR98 probe or a 145 nt antisenseGAM25 precursor transcript probe respectively. The Northern blot results of these experiments demonstrated processing of EST72223 RNA by Hela lysate (lanes 2-4, in Figs. 15B and 15C), into ~80bp and ~22bp segments, which reacted

with the MIR98 precursor probe (Fig. 15B), and into ~100bp and ~24bp segments, which reacted with the GAM25 precursor probe (Fig. 15C). These results demonstrate the processing of EST72223 by Hela lysate into MIR98 precursor and GAM25 precursor. It is also appreciated from Fig. 15C (lane 1) that Hela lysate itself reacted with the GAM25 precursor probe, in a number of bands, including a ~100bp band, indicating that GAM25-precursor is endogenously expressed in Hela cells. The presence of additional bands, higher than 100bp in lanes 5-9 probably corresponds to the presence of nucleotide sequences in Hela lysate, which contain the GAM25 sequence.

[0324] In addition, in order to demonstrate the kinetics and specificity of the processing of MIR98 and GAM25 precursors into their respective mature, "diced" segments, transcripts of MIR98 and of the bioinformatically predicted GAM25 precursors were similarly incubated with Hela S100 lysate, for 0 minutes, 30 minutes, 1 hour and 24 hours, and for 24 hours with the addition of EDTA, added to inhibit Dicer activity, following which RNA was harvested, run on a polyacrylamide gel and reacted with MIR98 and GAM25 precursor probes. Capped transcripts

were prepared for in vitro RNA cleavage assays with T7 RNA polymerase, including a m7G(5')ppp(5')G-capping reaction using the T7-mMessage mMachine kit (Ambion). Purified PCR products were used as template for the reaction. These were amplified for each assay with specific primers containing a T7 promoter at the 5' end and a T3 RNA polymerase promoter at the 3' end. Capped RNA transcripts were incubated at 30C in supplemented, dialysis concentrated, Hela S100 cytoplasmic extract (4C Biotech, Seneffe, Belgium). The Hela S100 was supplemented by dialysis to a final concentration of 20mM Hepes, 100mM KCl, 2.5mM MgCl2, 0.5mM DTT, 20% glycerol and protease inhibitor cocktail tablets (Complete mini Roche Molecular Biochemicals). After addition of all components, final concentrations were 100mM capped target RNA, 2mM ATP, 0.2mM GTP, 500U/ml RNasin, 25 microgram/ml creatine kinase, 25mM creatine phosphate, 2.5mM DTT and 50% S100 extract. Proteinase K, used to enhance Dicer activity (Zhang et al., EMBO J. 21, 5875-5885 (2002)) was dissolved in 50mM Tris-HCl pH 8, 5mM CaCl2, and 50% glycerol, was added to a final concentration of 0.6 mg/ml. Cleavage reactions were stopped by the addition of 8 volumes of proteinase K buffer

(200Mm Tris-Hcl, pH 7.5, 25m M EDTA, 300mM NaCl, and 2% SDS) and incubated at 65C for 15min at different time points (0, 0.5, 1, 4, 24h) and subjected to phenol/ chloroform extraction. Pellets were dissolved in water and kept frozen. Samples were analyzed on a segmented half 6%, half 13% polyacrylamide 1XTBE-7M Urea gel.

[0325] The Northern blot results of these experiments demonstrated an accumulation of a ~22bp segment which reacted with the MIR98 precursor probe, and of a ~24bp segment which reacted with the GAM25 precursor probe, over time (lanes 5-8). Absence of these segments when incubated with EDTA (lane 9), which is known to inhibit Dicer enzyme (Zhang et al., 2002), supports the notion that the processing of MIR98 and GAM25 precursors into their " diced" segments is mediated by Dicer enzyme, found in Hela lysate. Other RNases do not utilize divalent cations and are thus not inhibited by EDTA. The molecular sizes of EST72223, MIR-98 and GAM25 and their corresponding precursors are indicated by arrows.

[0326] Fig. 15D present Northern blot results of same above experiments with GAM25 probe (24 nt). The results clearly demonstrated the accumulation of mature GAM25 oligonucleotide after 24 h.

[0327] To validate the identity of the band shown by the lower arrow in figs. 15C and 15D, a RNA band parallel to a marker of 24 base was excised from the gel and cloned as in Elbashir et al (2001) and sequenced. Ninety clones corresponded to the sequence of mature GAM25 oligonucleotide, three corresponded to GAM25* (the opposite arm of the hairpin with a 1-3 nt 3' overhang) and two to the hairpin-loop.

[0328] GAM25 was also validated endogenously by sequencing from both sides from a HeLa YM100 total-RNA "ligation" libraries, utilizing hemispecific primers as described in Fig. 13.

[0329] Taken together, these results validate the presence and processing of a novel miRNA-like oligonucleotide, GAM25, which was predicted bioinformatically. The processing of this novel GAM oligonucleotide product, by Hela lysate from EST72223, through its precursor, to its final form was similar to that observed for known miRNA oligonucleotide, MIR98.

[0330] Transcript products were 705 nt (EST72223), 102 nt (MIR98 precursor), 125 nt (GAM25 precursor) long. EST72223 was PCR-amplified with T7-EST 72223 forward primer:

5'-TAATACGACTCACTATAGGCCCTTATTAGAGGATTCTGCT-3' and T3-EST72223 reverse primer:"-AATTAACCCTCACTAAAGGTTTTTTTTTCCTGAGACAGAGT-3'.MIR98 was PCR-amplified using EST72223 as a template with T7MIR98 forward primer: 5'-TAATACGACTCACTATAGGGTGAGGTAGTAAGTTGTATTGTT-3'and T3MIR98 reverse primer: 5'-AATTAACCCTCACTAAAGGGAAAGTAGTAAGTTGTATAGTT-3'. GAM25 was PCR-amplified using EST72223 as a template with GAM25 forward primer: 5'-GAGGCAGGAGAATTGCTTGA-3' and T3-EST72223 reverse primer:5'-AATTAACCCTCACTAAAGGCCTGAGACAGAGTCTTGCTC-3'.

[0331] It is appreciated that the data presented in Figs. 15A, 15B, 15C and 15D when taken together validate the function of the bioinformatic oligonucleotide detection engine 100 of Fig. 2. Fig. 15A shows a novel GAM oligonucleotide bioinformatically-detected by the bioinformatic oligonucleotide detection engine 100, and Figs. 15C and 15D show laboratory confirmation of the expression of this novel oligonucleotide. This is in accord with the engine training and validation methodology described hereinabove with

reference to Fig. 2.

[0332] Reference is now made to Figs. 16A-C, which schematically represent three methods that are employed to identify GAM FOLDED PRECURSOR RNA from libraries. Each method involves the design of specific primers for PCR amplification followed by sequencing. The libraries include hairpins as double-stranded DNA with two different adaptors ligated to their 5' and 3' ends.

[0333] Reference is now made to Fig. 16A, which depicts a first method that uses primers designed to the stems of the hairpins. Since the stem of the hairpins often has bulges, mismatches, as well as G-T pairing, which is less significant in DNA than is G-U pairing in the original RNA hairpin, the primer pairs were engineered to have the lowest possible match to the other strand of the stem. Thus, the F-Stem primer, derived from the 5' stem region of the hairpin, was chosen to have minimal match to the 3' stem region of the same hairpin. Similarly, the R-stem primer, derived from the 3' region of the hairpin (reverse complementary to its sequence), was chosen to have minimal match to the 5' stem region of the same hairpin. The F-Stem primer was extended in its 5' sequence with the T3 primer (5'-ATTAACCCTCACTAAAGGGA-3') and the R-

Stem primer was extended in its 5' sequence with the T7 primer (5'- TAATACGACTCACTATAGGG). The extension is needed to obtain a large enough fragment for direct sequencing of the PCR product. Sequence data from the amplified hairpins is obtained in two ways. One way is the direct sequencing of the PCR products using the T3 primer that matches the extension of the F-Stem primer. Another way is the cloning of the PCR products into a plasmid, followed by PCR screening of individual bacterial colonies using a primer specific to the plasmid vector and either the R-Loop (Fig. 16B) or the F-Loop (Fig. 16C) primer. Positive PCR products are then sent for direct sequencing using the vector-specific primer.

[0334] Reference is now made to Fig. 16B, which depicts a second method in which R-Stem primer and R-Loop primers are used in a nested-PCR approach. First, PCR is performed with the R-Stem primer and the primer that matches the 5' adaptor sequence (5-ad primer). PCR products are then amplified in a second PCR using the R-Loop and 5-ad primers. As mentioned hereinabove, sequence data from the amplified hairpins is obtained in two ways. One way is the direct sequencing of the PCR products using the 5-ad primer. Another way is the cloning of

the PCR products into a plasmid, followed by PCR screening of individual bacterial colonies using a primer specific to the plasmid vector and F-Stem primer. Positive PCR products are then sent for direct sequencing using the vector-specific primer. It should be noted that optionally an extended R-Loop primer is designed that includes a T7 sequence extension, as described hereinabove (Fig. 16A) for the R-Stem primer. This is important in the first sequencing option in cases where the PCR product is too short for sequencing.

[0335] Reference is now made to Fig. 16C, which depicts a third method, which is the exact reverse of the second method described hereinabove (Fig. 16B). F-Stem and F-Loop primers are used in a nested-PCR approach. First, PCR is performed with the F-Stem primer and the primer that matches the 3' adaptor sequence (3-ad primer). PCR products are then amplified in a second PCR using the F-Loop and 3-ad primers. As in the other two methods, sequence data from the amplified hairpins is obtained in two ways. One way is the direct sequencing of the PCR products using the F-Loop primer. Another way is the cloning of the PCR products into a plasmid, followed by PCR screening of individual bacterial colonies using a primer

specific to the plasmid vector and R-Stem primer. Positive PCR products are then sent for direct sequencing using the vector-specific primer. It should be noted that optionally an extended F-Loop primer is designed that includes a T3 sequence extension, as described hereinabove (Fig. 16A) for the F-Stem primer. This is important in the first sequencing option in cases where the PCR product is too short for sequencing and also in order to enable the use of T3 primer.

[0336] In an embodiment of the present invention, the three methods mentioned hereinabove may be employed to validate the expression of GAM FOLDED PRECURSOR RNA.

[0337] Reference is now made to Fig. 17A, which is a flow chart with a general description of the design of the microarray to identify expression of published miRNA oligonucleotides, and of novel GAM oligonucleotides of the present invention.

[0338] A microarray that identifies miRNA oligonucleotides is designed (Fig. 17B). The DNA microarray is prepared by Agilent according to their SurePrint Procedure (reference describing their technology can be obtained from the Agilent website, http://www.agilent.com). In this procedure, the oligonucleotide probes are synthesized on the glass sur-

face. Other methods can also be used to prepare such microarray including the printing of pre-synthesized oligonucleotides on glass surface or using the photolithography method developed by Affymetrix (Lockhart DJ et al., Nat Biotechnol.14:1675-1680 (1996)). The 60-mer sequences from the design are synthesized on the DNA microarray. The oligonucleotides on the microarray, termed "probes" are of the exact sequence as the designed 60-mer sequences. Importantly, the 60-mer sequences and the probes are in the sense orientation with regards to the miRNA oligonucleotides. Next, a cDNA library is created from size-fractionated RNA, amplified, and converted back to RNA (Fig. 17C). The resulting RNA is termed "cRNA". The conversion to RNA is done using a T7 RNA polymerase promoter found on the 3' adaptor (Fig. 17C; T7 NcoI-RNA-DNA 3'Adaptor). Since the conversion to cRNA is done in the reverse direction compared to the orientation of the miRNA oligonucleotides, the cRNA is reverse complementary to the probes and is able to hybridize to it. This amplified RNA is hybridized with the microarray that identifies miRNA oligonucleotides, and the results are analyzed to indicate the relative level of miRNA oligonucleotides (and hairpins) that are present in

the total RNA of the tissue (Fig. 18).

[0339] Reference is now made to Fig. 17B, which describes how the microarray to identify miRNA oligonucleotides is designed. miRNA oligonucleotide sequences or potential predicted miRNA oligonucleotides are generated by using known or predicted hairpins as input. Overlapping potential miRNA oligonucleotides are combined to form one larger sub-sequence within a hairpin.

[0340] To generate non-expressed sequences (tails), artificial sequences are generated that are 40 nts in length, which do not appear in the respective organism genome, do not have greater than 40% homology to sequences that appear in the genome, and with no 15-nucleotide window that has greater than 80% homology to sequences that appear in the genome.

[0341] To generate probe sequences, the most probable miRNA oligonucleotide sequences are placed at position 3 (from the 5' end) of the probe. Then, a tail sub-sequence to the miRNA oligonucleotide sequence was attached such that the combined sequence length will meet the required probe length (60 nts for Agilent microarrays).

[0342] The tails method provides better specificity compared to the triplet method. In the triplet method, it cannot be as-

certained that the design sequence, and not an uncontrolled window from the triplet probe sequence, was responsible for hybridizing to the probe. Further, the tails method allows the use of different lengths for the potential predicted miRNA oligonucleotide (of combined, overlapping miRNA oligonucleotides).

[0343] Hundreds of control probes were examined in order to ensure the specificity of the microarray. Negative controls contain probes which should have low intensity signal. For other control groups, the concentration of certain specific groups of interest in the library are monitored. Negative controls include tail sequences and non-hairpin sequences. Other controls include mRNA for coding genes, tRNA, and snoRNA.

[0344] For each probe that represents known or predicted miRNA oligonucleotides, additional mismatch probes were assigned in order to verify that the probe intensity is due to perfect match (or as close as possible to a perfect match) binding between the target miRNA oligonucleotide cRNA and its respective complementary sequence on the probe. Mismatches are generated by changing nucleotides in different positions on the probe with their respective complementary nucleotides (A <-> T, G <-> C, and vice

versa). Mismatches in the tail region should not generate a significant change in the intensity of the probe signal, while mismatches in the miRNA oligonucleotide sequences should induce a drastic decrease in the probe intensity signal. Mismatches at various positions within the miRNA oligonucleotide sequence enable us to detect whether the binding of the probe is a result of perfect match or, alternatively, nearly perfect match binding.

[0345] Based on the above scheme, we designed a DNA microarray prepared by Agilent using their SurePrint technology. Table 11 is a detailed list of microarray chip probes

[0346] KNOWN miRNA OLIGONUCLEOTIDES:

[0347] The miRNA oligonucleotides and their respective precursor sequences are taken from Sanger Database to yield a total of 186 distinct miRNA oligonucleotide and precursor pairs. The following different probes are constructed:

[0348] 1. SINGLE miRNA OLIGONUCLEOTIDE PROBES:

[0349] From each precursor, 26-mer containing the miRNA oligonucleotide were taken, then assigned 3 probes for each extended miRNA oligonucleotide sequence: 1. the 26-mer are at the 5' of the 60-mer probe, 2. the 26-mer are at the 3' of the 60-mer probe, 3. the 26-mer are in

the middle of the 60-mer probe. Two different 34-mer subsequences from the design tails are attached to the 26-mer to accomplish 60-mer probe. For a subset of 32 of Single miRNA oligonucleotide probes, six additional mismatches mutations probes were designed:

[0350] 4 block mismatches at 5' end of the miRNA oligonucleotide;

[0351] 6 block mismatches at 3' end of the miRNA oligonucleotide;

[0352] 1 mismatch at position 10 of the miRNA oligonucleotide;

[0353] 2 mismatches at positions 8 and 17 of the miRNA oligonucleotide;

[0354] 3 mismatches at positions 6, 12 and 18 of the miRNA oligonucleotide; and

[0355] 6 mismatches at different positions out of the miRNA oligonucleotide.

[0356] 2. DUPLEX miRNA OLIGONUCLEOTIDE PROBES:

[0357] From each precursor, a 30-mer containing the miRNA oligonucleotide was taken, then duplicated to obtain 60-mer probe. For a subset of 32 of probes, three additional mismatch mutation probes were designed:

[0358] 2 mismatches on the first miRNA oligonucleotide;

[0359] 2 mismatches on the second miRNA oligonucleotide; and

[0360] 2 mismatches on each of the miRNA oligonucleotides.

[0361] 3. TRIPLET miRNA OLIGONUCLEOTIDE PROBES:

[0362] Following Krichevsky's work (Krichevsky et al., RNA 9:1274-1281 (2003)), head to tail ~22-mer length miRNA oligonucleotide sequences were attached to obtain 60-mer probes containing up to three repeats of the same miRNA oligonucleotide sequence. For a subset of 32 probes, three additional mismatch mutation probes were designed:

[0363] 2 mismatches on the first miRNA oligonucleotide;

[0364] 2 mismatches on the second miRNA oligonucleotide; and

[0365] 2 mismatches on each of the miRNA oligonucleotides.

[0366] 4. PRECURSOR WITH miRNA OLIGONUCLEOTIDE PROBES:

[0367] For each precursor, 60-mer containing the miRNA oligonucleotide were taken.

[0368] 5. PRECURSOR WITHOUT miRNA OLIGONUCLEOTIDE PROBES:

[0369] For each precursor, a 60-mer containing no more then 16-mer of the miRNA oligonucleotide was taken. For a subset of 32 probes, additional mismatch probes contain-

ing four mismatches were designed.

[0370] CONTROL GROUPS:

[0371] 1. 100 60-mer sequences from representative ribosomal RNAs.

[0372] 2. 85 60-mer sequences from representatives tRNAs.

[0373] 3. 19 60-mer sequences from representative snoRNA.

[0374] 4. 294 random 26-mer sequences from human genome not contained in published or predicted precursor sequences, placing them at the probe's 5' and attached 34-mer tail described above.

[0375] 5. Negative Control: 182 different 60-mer probes contained different combinations of 10 nt-long sequences, in which each 10 nt-long sequence is very rare in the human genome, and the 60-mer combination is extremely rare.

[0376] PREDICTED GAM RNAs:

[0377] There are 8642 pairs of predicted GAM RNA and their respective precursors. From each precursor, a 26-mer containing the GAM RNA was placed at the 5' of the 60-mer probe and a 34-mer tail was attached to it. For each predicted probe, a mutation probes with 2 mismatches at positions 10 and 15 of the GAM RNA were added.

[0378] For a subset of 661 predicted precursors, up to 2 probes

each containing one side of the precursor including any possible GAM RNA in it were added.

[0379] Microarray analysis:

[0380] Based on known miRNA oligonucleotide probes, a preferred position of the miRNA oligonucleotide on the probe was evaluated, and hybridization conditions adjusted and the amount of cRNA to optimize microarray sensitivity and specificity ascertained. Negative controls are used to calculate background signal mean and standard deviation. Different probes of the same miRNA oligonucleotide are used to calculate signal standard deviation as a function of the signal.

[0381] For each probe, BG_Z_Score = (log(probe signal) - mean of log(negative control signal))/(log(negative control signal) standard deviation) were calculated.

[0382] For a probe with a reference probe with 2 mismatches on the miRNA oligonucleotide, MM_Z_Score MM_Z_Score = (log(perfect match signal) - log(reference mismatch signal))/(standard deviation of log(signals) as the reference mismatch log(signal)) were calculated.

[0383] BG_Z_Score and MM_Z_Score are used to decide whether the probe is on and its reliability.

[0384] Reference is now made to Fig. 17C, which is a flowchart

describing how the cDNA library was prepared from RNA and amplified. The general procedure was performed as described previously (Elbashir SM, Lendeckel W, Tuschl T. RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Dev. 2001 15:188-200) with several modifications, which will be described hereinbelow.

[0385] First, the starting material is prepared. Instead of starting with standard total RNA, the total RNA was size-fractionated using an YM-100 Microcon column (Millipore Corporation, Billerica, Massachusetts, USA) in the present protocol. Further, the present protocol uses human tissue or cell lines instead of a Drosophila in vitro system as starting materials. Finally, 3 micrograms of size-fractionated total RNA was used for the ligation of adaptor sequences.

[0386] Libraries used for microarray hybridization are listed hereinbelow: "A" library is composed of a mix of libraries from Total HeLa YM100 RNA and Nuclear HeLa YM100 RNA; "B" library is composed of a mix of libraries from Total HEK293 YM100 RNA and Nuclear HEK293 YM100 RNA; "C" library is composed of a mix of YM100 RNA libraries from Total PC3, Nuclear PC3 and from PC3 cells in which Dicer expression was transiently silenced by Dicer

specific siRNA; "D" library is prepared from YM100 RNA from Total Human Brain (Ambion Cat#7962); "E" library is prepared from YM100 RNA from Total Human Liver (Ambion Cat#7960); "F" library is prepared from YM100 RNA from Total Human Thymus (Ambion Cat#7964); "G" library is prepared from YM100 RNA from Total Human Testis (Ambion Cat#7972); and "H" library is prepared from YM100 RNA from Total Human Placenta (Ambion Cat#7950).

[0387] Library letters appended by a numeral "1" or "2" are digested by XbaI (NEB); Library letters affixed by a numeral "3" are digested by Xba1 and SpeI (NEB); Library letters appended by a numeral "4" are digested by Xba1 and the transcribed cRNA is then size-fractionated by YM30, retaining the upper fraction consisting of 60 nts and longer; Library letters affixed by a numeral "5" are digested by Xba1 and the transcribed cRNA is then size-fractionated by YM30 retaining the flow-through fraction consequently concentrated with YM10 consisting of 30 nts-60 nts; Library letters affixed by a numeral "6" are digested by Xba1 and the DNA is fractionated on a 13% native acrylamide gel from 40-60 nt, electroeluted on a GeBaFlex Maxi column (GeBa Israel), and lyophilized; Library letters affixed

by a numeral "7" are digested by Xba1 and the DNA is fractionated on a 13% native acrylamide gel from 80-160 nt, electroeluted and lyophilized.

[0388] Next, unique RNA-DNA hybrid adaptor sequences with a T7 promoter were designed. This step is also different than other protocols that create libraries for microarrays. Most protocols use complements to the polyA tails of mRNA with a T7 promoter to amplify only mRNA. However, in the present invention, adaptors are used to amplify all of the RNA within the size-fractionated starting material. The adaptor sequences are ligated to the size-fractionated RNA as described in Fig. 13, with subsequent gel-fractionation steps. The RNA is then converted to first strand cDNA using reverse transcription.

[0389] Next, the cDNA is amplified using PCR with adaptor-specific primers. At this point, there is the optional step of removing the tRNA, which is likely to be present because of its low molecular weight, but may add background noise in the present experiments. All tRNA contain the sequence ACC at their 3' end, and the adaptor contains GGT at its 5' end. This sequence together (GGTACC) is the target site for NcoI restriction digestion. Thus, adding the restriction enzyme NcoI either before or during PCR am-

plification will effectively prevent the exponential amplification of the cDNA sequences that are complements of the tRNAs.

[0390] The amplified DNA is restriction enzyme-digested with Xba1 (and, optionally, with Pst or SpeI) to remove the majority of the adaptor sequences that were initially added to the RNA. Using the first set of RNA-DNA hybrid adaptors listed below, the first two sets of primers listed below, and Xba1 restriction digest yields the following cRNA products: 5'GGCCA - PRE/miRNA- UAUCUAG, where PRE is defined as GAM PRECURSOR (palindrome). Using the second set of RNA-DNA hybrid adaptors listed below, the second set of primers listed below, and Xba1and Pst restriction digest yields the following, smaller cRNA products: 5'GG-PRE/miRNA - C*.

[0391] Then, cDNA is transcribed to cRNA utilizing an RNA polymerase e.g. T7 dictated by the promoter incorporated in the adaptor. cRNA may be labeled in the course of transcription with aminoallyl or fluorescent nucleotides such as Cy3- or Cy5-UTP and CTP among other labels, and cRNA sequences thus transcribed and labeled are hybridized with the microarray.

[0392] The following RNA-DNA hybrid adaptors are included in

the present invention:

[0393] Name: T7 NcoI-RNA-DNA 3'Adapter

[0394] Sequence:
5'(5phos)rUrGrGCCTATAGTGAGTCGTATTA(3InvdT)3'

[0395] 2. Name: 5Ada RNA-DNA XbaBseRI

[0396] Sequence: 5' AAAGGAGGAGCTCTAGrArUrA 3' or optionally:

[0397] 3. Name: 5Ada MC RNA-DNA PstAtaBser

[0398] Sequence: 5' CCTAGGAGGAGGACGTCTGrCrArG 3'

[0399] 4. Name: 3'Ada nT7 MC RNA-DNA

[0400] Sequence: 5' (5phos) rCrCrUATAGTGAGTCGTATTATCT (3InvdT)3'

[0401] The following DNA primers are included in the present invention:

[0402] 1. Name: T7 NcoI-RT-PCR primer

[0403] Sequence: 5' TAATACGACTCACTATAGGCCA 3'

[0404] 2. Name: T7NheI SpeI-RT-PCR primer

[0405] Sequence: 5' GCTAGCACTAGTTAATACGACTCACTATAGGCCA 3'

[0406] 3. Name: 5Ada XbaBseRI Fwd

[0407] Sequence: 5' AAAGGAGGAGCTCTAGATA 3'

[0408] 4. Name: Pst-5Ada XbaBseRI Fwd

[0409] Sequence: 5' TGACCTGCAGAAAGGAGGAGCTCTAGATA 3'

[0410] or optionally:

[0411] 5. Name: 5Ada MC PstAtaBser fwd

[0412] Sequence: 5' ATCCTAGGAGGAGGACGTCTGCAG 3'

[0413] 6. Name: RT nT7 MC XbaI

[0414] Sequence: 5' GCTCTAGGATAATACGACTCACTATAGG 3'

[0415] Reference is now made to Fig. 18A, which demonstrates the detection of known miRNA oligonucleotides and of novel GAM oligonucleotides, using a microarray constructed and operative in accordance with a preferred embodiment of the present invention. Based on negative control probe intensity signals, we evaluated the background, non-specific, logarithmic intensity distribution, and extracted its mean, designated BG_mean, and standard deviation, designated BG_std. In order to normalize intensity signals between different microarray experiments, a Z score, which is a statistical measure that quantifies the distance (measured in standard deviations) that

a data point is from the mean of a data set, was calculated for each probe with respect to the negative control using the following Z score formula: Z = (logarithm of probe signal BG_mean)/BG_std. We performed microarray experiments using RNA extracted from several different tissues and we calculated each probes maximum Z score. Fig. 18A shows the percentages of known, predicted and negative control groups that have a higher max Z score than a specified threshold as a function of max Z score threshold. The negative control group plot, included as a reference, considers probe with a max Z score greater then 4 as a reliable probe with meaningful signals. The sensitivity of our method was demonstrated by the detection of almost 80% of the known published miRNA oligonucleotides in at least one of the examined tissues. At a threshold of 4 for the max Z score, 28% of the predicted GAMs are present in at least one of the examined tissues.

[0416] Reference is now made to Fig. 18B, which is a line graph showing specificity of hybridization of a microarray constructed and operative in accordance with a preferred embodiment of the present invention and described hereinabove with reference to Figs. 17A-17C.

[0417] The average signal of known miRNA oligonucleotides in

Library A2 is presented on a logarithmic scale as a function of the following probe types under two different hybridization conditions: 50 C and 60 C: perfect match (PM), six mismatches on the tail (TAIL MM), one mismatch on the miRNA oligonucleotide (1MM), two separate mismatches on the miRNA oligonucleotide (2MM), three separate mismatches on the miRNA oligonucleotide (3MM). The relative equality of perfect match probes and probes with the same miRNA oligonucleotide but many mismatches over the tail attest to the independence between the tail and the probe signal. At a hybridization temperature of 60 C, one mismatch in the middle of the miRNA oligonucleotide is enough to dramatically reduce the probe signal. Conducting chip hybridization at 60 C ensures that a probe has a very high specificity.

[0418] It is appreciated that these results demonstrate the specificity of the microarray of the present invention in detecting expression of miRNA oligonucleotides.

[0419] Reference is now made to Fig. 18C, which is a summary table demonstrating detection of known miRNA oligonucleotides using a microarray constructed and operative in accordance with a preferred embodiment of the present invention and described hereinabove with reference to

Figs. 17A-17C.

[0420] Labeled cRNA from HeLa cells and Human Liver, Brain, Thymus, Placenta, and Testes was used for 6 different hybridizations. The table contains the quantitative values obtained for each miRNA oligonucleotide probe. For each miRNA oligonucleotide, the highest value (or values) is given in bolded font while lower values are given in regular font size. Results for MIR-124A, MIR-9 and MIR-122A are exactly as expected from previous studies. The References column contains the relevant references in the published literature for each case. In addition to these miRNA oligonucleotides, the table shows other known miRNA oligonucleotides that are expressed in a tissue-specific manner. The results indicate that MIR-128A, MIR-129 and MIR-128B are highly enriched in Brain; MIR-194, MIR-148 and MIR-192 are highly enriched in Liver; mIR-96, MIR-150, MIR-205, MIR-182 and MIR-183 are highly enriched in Thymus; MIR-204, MIR-10B, MIR-154 and MIR134 are highly enriched in Testes; and MIR-122, MIR-210, MIR-221, MIR-141, MIR-23A, MIR-200C and MIR-136 are highly enriched in Placenta. In most cases, low but significant levels are observed in the other tissues. However, in some cases, miRNA oligonucleotides are also expressed at

relative high levels in an additional tissue.

[0421] It is appreciated that these results reproduce previously published studies of expression of known miRNA oligonucleotides. These results demonstrate the reliability of the microarray of the present invention in detecting expression of published miRNA oligonucleotides, and of novel GAM oligonucleotides of the present invention.

DETAILED DESCRIPTION OF TABLES

[0422] Table 1 comprises data relating the SEQ ID NO of oligonucleotides of the present invention to their corresponding GAM NAME, and contains the following fields: GAM SEQ-ID: GAM SEQ ID NO, as in the Sequence Listing; GAM NAME: Rosetta Genomics Ltd. nomenclature (see below); GAM RNA SEQUENCE: Sequence (5' to 3') of the mature, "diced" GAM RNA; GAM ORGANISM: identity of the organism encoding the GAM oligonucleotide; GAM POS: Dicer-cut location (see below); and

[0423] Table 2 comprises detailed textual description according to the description of Fig. 1 of each of a plurality of novel GAM oligonucleotides of the present invention, and contains the following fields: GAM NAME: Rosetta Genomics Ltd. nomenclature (see below); GAM ORGANISM: identity of the organism encoding the GAM oligonucleotide; PRE-

CUR SEQ-ID:GAM precursor Seq-ID, as in the Sequence Listing; PRECURSOR SEQUENCE: Sequence (5' to 3') of the GAM precursor; GAM DESCRIPTION: Detailed description of GAM oligonucleotide with reference to Fig. 1; and

[0424] Table 3 comprises data relating to the source and location of novel GAM oligonucleotides of the present invention, and contains the following fields: GAM NAME: Rosetta Genomics Ltd. nomenclature (see below); PRECUR SEQ-ID: GAM precursor SEQ ID NO, as in the Sequence Listing; GAM ORGANISM: identity of the organism encodes the GAM oligonucleotide; SOURCE: For human GAM- chromosome encoding the human GAM oligonucleotide, otherwise- accession ID (GenBank, NCBI); STRAND: Orientation of the strand, "+" for the plus strand, "-" for the minus strand; SRC-START OFFSET: Start offset of GAM precursor sequence relative to the SOURCE; SRC-END OFFSET: End offset of GAM precursor sequence relative to the SOURCE; and

[0425] Table 4 comprises data relating to GAM precursors of novel GAM oligonucleotides of the present invention, and contains the following fields: GAM NAME: Rosetta Genomics Ltd. nomenclature (see below); PRECUR SEQ-ID: GAM precursor Seq-ID, as in the Sequence Listing; GAM

ORGANISM: identity of the organism encoding the GAM oligonucleotide; PRECURSOR-SEQUENCE: GAM precursor nucleotide sequence (5' to 3'); GAM FOLDED PRECURSOR RNA: Schematic representation of the GAM folded precursor, beginning 5' end (beginning of upper row) to 3' end (beginning of lower row), where the hairpin loop is positioned at the right part of the draw; and

[0426] Table 5 comprises data relating to GAM oligonucleotides of the present invention, and contains the following fields: GAM NAME: Rosetta Genomics Ltd. nomenclature (see below); GAM ORGANISM: identity of the organism encoding the GAM oligonucleotide; GAM RNA SEQUENCE: Sequence (5' to 3') of the mature, "diced" GAM RNA; PRECUR SEQ-ID: GAM precursor Seq-ID, as in the Sequence Listing; GAM POS: Dicer-cut location (see below); and

[0427] Table 6 comprises data relating SEQ ID NO of the GAM target gene binding site sequence to TARGET gene name and target binding site sequence, and contains the following fields: TARGET BINDING SITE SEQ-ID: Target binding site SEQ ID NO, as in the Sequence Listing; TARGET ORGANISM: identity of organism encode the TARGET gene; TARGET: GAM target gene name; TARGET BINDING SITE SEQUENCE: Nucleotide sequence (5' to 3') of the target

binding site; and

[0428] Table 7 comprises data relating to target-genes and binding sites of GAM oligonucleotides of the present invention, and contains the following fields: GAM NAME: Rosetta Genomics Ltd. nomenclature (see below); GAM ORGANISM: identity of the organism encoding the GAM oligonucleotide; GAM RNA SEQUENCE: Sequence (5' to 3') of the mature, "diced" GAM RNA; TARGET: GAM target gene name; TARGET REF-ID: For human target genes-Target accession number (RefSeq, GenBank); Otherwise-the location of the target gene on the genome annotation. TARGET ORGANISM: identity of organism encode the TARGET gene; UTR: Untranslated region of binding site/s (3' or 5'); TARGET BS-SEQ: Nucleotide sequence (5' to 3') of the target binding site; BINDING SITE-DRAW: Schematic representation of the binding site, upper row represent 5' to 3' sequence of the TARGET, Lower row represent 3' to 5' Sequence of the GAM RNA; GAM POS: Dicer-cut location (see below); and

[0429] Table 8 comprises data relating to functions and utilities of novel GAM oligonucleotides of the present invention, and contains the following fields: GAM NAME: Rosetta Genomics Ltd. nomenclature (see below); GAM RNA SE-

QUENCE: Sequence (5' to 3') of the mature, "diced" GAM RNA; GAM ORGANISM: identity of the organism encoding the GAM oligonucleotide; TARGET: GAM target gene name; TARGET ORGANISM: identity of organism encode the TARGET gene; GAM FUNCTION: Description of the GAM functions and utilities; GAM POS: Dicer-cut location (see below); and

[0430] Table 9 comprises references of GAMs target genes and contains the following fields: TARGET: Target gene name; TARGET ORGANISM: identity of organism encode the TARGET gene; REFERENCES: reference relating to the target gene; and

[0431] Table 10 comprises data relating to novel GR (Genomic Record) polynucleotides of the present invention, and contains the following fields: GR NAME: Rosetta Genomics Ltd. nomenclature (see below); GR ORGANISM: identity of the organism encoding the GR polynucleotide; GR DESCRIPTION: Detailed description of a GR polynucleotide, with reference to Fig. 9; and

[0432] Table 11 comprises data of all sequences printed on the microarray of the microarray experiment, as described herein above with reference to Fig. 17 and include the following fields: PROBE SEQUENCE: the sequence that was

printed on the chip PROBE TYPE: as described in detail in Fig. 17 in chip design section and summarized as follows: Known: published miRNA sequence; Known_mis1: similar to published miRNA sequence, but with 1 mismatch mutation on the miRNA sequence; Known_mis2: similar to published miRNA sequence, but with 2 mismatch mutations on the miRNA sequence; Known_mis3: similar to published miRNA sequence, but with 3 mismatch mutations on the miRNA sequence; Known_mis4: similar to published miRNA sequence, but with 6 mismatch mutations on regions other than the miRNA sequence; Predicted: predicted GAM RNA sequences; Mismatch: sequences that are similar to predicted GAM RNA sequences but with 2 mismatches; Edges1: left half of GAM RNA sequences; Edges2: right half of GAM RNA sequences extended with its hairpin precursor (palindrome); Control1: negative control; Control2: random sequences; Control3: tRNA; Control4: snoRNA; Control5: mRNA; Control6: other; GAM RNA SEQ ID/MIR NAME: GAM oligonucleotide using Rosetta Genomics Ltd. Nomenclature (see below) or published miRNA oligonucleotide terminology; GAM RNA SEQUENCE: Sequence (5' to 3') of the mature, "diced" GAM RNA; LIBRARY: the library name as defined in Fig. 17C;

SIGNAL: Raw signal data for library; BACKGROUND Z-SCORE: Z-score of probe signal with respect to background, negative control signals; MISMATCH Z-SCORE: Z-score of probe signal with respect to its mismatch probe signal; and

[0433] Table 12 comprises data related to the GAM RNA SEQUENCEs included in the present invention that were validated by laboratory means. If the validated sequence appeared in more than one GAM precursor, the GAM RNA SEQ-ID indicated may be arbitrarily chosen. The table includes the following fields: VALIDATION METHOD: the type of validation performed on the sequence. The microrray validations are divided into four groups: a) "Chip strong" refers to GAM oligonucleotide sequences whose intensity (SIGNAL) on the microarray "chip" was more than 6 standard deviations above the background intensity, and the differential to the corresponding mismatch intensity was more than 2 standard deviations, where in this case the standard deviation is of the intensity of identical probes; b) "Chip" refers to GAM oligonucleotide sequences, whose intensity was more than 4 standard deviations above the background intensity; c) "Sequenced" refers to GAM oligonucleotide sequences that were se-

quenced; and d) "Chip strong, Sequenced" refers to miRNA oligonucleotide sequences that were both detected in the microarray as "Chip strong" and sequenced. "Sequenced" is described hereinabove with reference to Fig. 13. Other validations are from microarray experiments as described hereinabove with reference to Figs. 17A-C and 18A-C; SIGNAL: a raw signal data; BACKGROUND Z-SCORE: a Z-score of probe signal with respect to background, negative control signals; MISMATCH Z-SCORE: a Z-score of probe signal with respect to its mismatch probe signal; and

[0434] Table 13 comprises sequence data of GAMs associated with different bacterial infections. Each row refers to a specific bacterial infection, and lists the SEQ ID NOs of GAMs that target genes associated with that bacterial infection. The table contains the following fields: ROW#: index of the row number; INFECTION NAME: name of the infecting organism; and SEQ ID NOs OF GAMS ASSOCIATED WITH INFECTION: list of sequence listing IDs of GAMs targeting genes that are associated with the specified infection.

[0435] The following conventions and abbreviations are used in the tables: The nucleotide "U" is represented as "T" in the

tables, and;

[0436] GAM NAME or GR NAME are names for nucleotide sequences of the present invention given by RosettaGenomics Ltd. nomenclature method. All GAMs/GRs are designated by GAMx/GRx where x is a unique ID.

[0437] GAM POS is a position of the GAM RNA on the GAM PRECURSOR RNA sequence. This position is the Dicer-cut location: A indicates a probable Dicer-cut location; B indicates an alternative Dicer-cut location.

[0438] All human nucleotide sequences of the present invention as well as their chromosomal location and strand orientation are derived from sequence records of UCSC-hg16 version, which is based on NCBI, Build34 database (April, 2003).

[0439] All bacterial sequences of the present invention as well as their genomic location are derived from NCBI, RefSeq database.